Medical Physics

INTRODUCTION

The broadest concept that includes everything around us and our body is the matter. It is impossible to give a usual logical definition of matter, in which a broader concept is given, and then a sign of the object of definition is stated since no broader concept than matter. Therefore, instead of giving definition, it is often said that the matter was the objective reality given us in sensations.

The matter does not exist without impermanence. The impermanence is understood to be all the changes and processes taking place in the Universe. Conventionally, different and four types can represent various forms of the impermanence: physical, chemical, biological, and social. This makes it possible to classify different sciences depending on what type of impermanence they are studying. Physics studies the physical form of matter motion.

In more detail, the physical form of impermanence can be subdivided into mechanical, molecular-thermal, electromagnetic, nuclear, and intra-nuclear ones. Surely, this division is agreed. Nevertheless, physics as an academic discipline is usually represented by the foregoing sections.

Physics, like other sciences, uses different methods of research, but all they correspond ultimately to the unity of theory and practice and reflect the general scientific approach to cognition of the surrounding reality: observation, thinking and experiment. On the basis of observations, theories are created, laws and hypotheses are formulated; after that they are tested and used in practice. Practice is a criterion of theories, it makes it possible to define them more precisely. New theories and laws are formulated; they are tested by practice again. In this way, a person advances to more perfect understanding of the world around him or her.

Various forms of impermanence are interdependent and interrelated; that fact causes origin of new sciences, lying at the junction of the former ones – biophysics, astrophysics, chemical physics, etc., as well as use of the achievements of one science for development of another one.

The reader is naturally interested in connection between physics and medicine. Penetration of physical knowledge, techniques and instruments into medicine is quite multi-faceted; below there are just some of the aspects of this connection.

Physical processes in a body. Biophysics

Despite the complexity and interrelation of various processes in a human body, it is often possible to identify processes close to physical ones. E.g., such a complex physiological process as blood circulation is basically physical since it is connected with liquid flow (hydrodynamics), spread of elastic vibrations through vessels (vibrations and waves), mechanical work of the heart (mechanics), generation of bioelectric potentials (electricity), etc. Respiration is connected with motion of gas (aerodynamics), heat transfer (thermodynamics), evaporation (phase transformations), etc.

In the body, in addition to physical macroprocesses occurring in inanimate nature, there are molecular processes that ultimately determine the behavior of biological systems. Understanding the physics of such microprocesses is necessary for a correct assessment of the state of the body, the nature of certain diseases, action of medications, etc.

In solving all these problems, physics is so connected with biology that an independent science of biophysics is generated, which studies physical and physicochemical processes in living body, as well as the ultrastructure of biological systems at all levels of their organization — from submolecular and molecular one to a cell and the whole body.

Physical methods of diagnosis and research of biological systems

Many methods of diagnostics and research are based on using physical principles and ideas. Most modern devices for medical application are physical devices according to their design. To illustrate this, it is sufficient to consider some examples within information known to the reader from the course of secondary school.

Being a mechanical quantity the blood pressure is an index used to assess a number of conditions. Listening to sounds, whose sources are inside the body, yields information about normal or pathological functioning of organs. A clinical thermometer, whose work is based on thermal expansion of mercury, is a quite common diagnostic device. For the past decade, thanks to development of electronic devices, a diagnostic method based on recording biopotentials generated in a living body, has come into wide use. A method of electrocardiography is the most known; it consists in recording biopotentials reflecting cardiac activity. The role of a microscope for biomedical research is common known. Modern medical devices based on fiber optics make it possible to examine internal cavities of a body.

Spectral analysis is used in forensic medicine, hygiene, pharmacology and biology; achievements of atomic and nuclear physics are used for quite known diagnostic methods: X-ray diagnostics and isotopic labeling.

Therapeutic exposure to physical factors

In the set of different treatment methods, physical factors play a certain role. Here are some of them. Plaster bandage for fractures is a mechanical position lock for damaged organs. Cooling (ice) and heating (hot-water bag) for treatment are based on thermal effects. Electrical and electromagnetic controls are widely used in physiotherapy. Visible and invisible light (ultraviolet and infrared radiation), X-rays, and gamma-radiation are used for therapeutic purposes.

Physical properties of materials used in medicine. Physical properties of biological systems

Bandages, instruments, electrodes, prosthetic appliances, etc. used in medicine work in conditions of environmental impact, also immediately inside biological media. To estimate the applicability of such products in real conditions, it is necessary to have information about physical properties of materials from which they are made. Thus in manufacture of prosthetic appliances (false teeth, vessels, valves, etc.) it is essential to know mechanical strength, resistance to multiple loads, elasticity, thermal conductivity, electrical conductivity, and other physical properties.

In some cases, it is important to know physical properties of biological systems in order to estimate their viability or ability to withstand certain external impacts. It is possible to diagnose some diseases monitoring changes of physical properties of biological objects.

Physical properties and characteristics of the environment

A living body functions normally only interacting with the environment. It reacts dramatically to changes in such physical characteristics of the environment as temperature, humidity, air pressure, etc. The impact of the environment on a body is taken into account not only as an external factor: it can be used for therapeutic purposes (climatotherapy and pneumatotherapy). These examples confirm that a physician should be able to estimate physical properties and characteristics of the environment.

The foregoing examples of application of physics in medicine constitute *medical physics*, which is a set of sections of applied physics and biophysics that deal with physical laws, phenomena, processes, and characteristics in the context of solving medical problems.

Medicine and technical devices

Modern medicine is based on wide use of various equipment, which is mostly physical by its design, so in the course of medical and biological physics, the structure and principles of operation of basic medical equipment are considered.

Medicine, computers and mathematics

Computers are most widely used for both processing medical research findings and diagnosing of a disease. In addition, mathematics is used to describe processes occurring in living systems, as well as to develop and analyze corresponding models. Mathematical statistics is used to identify types of diseases, calculate prevalence of epidemics and for other purposes.

БИБЛИОТЕКА

the structure and principle of operation of [illegible] [illegible]

Medicine, computers and mathematics

Computers are most widely used for both processing medical research findings and diagnosing of a disease. In addition, mathematics is used to describe processes occurring [illegible] and to [illegible] corresponding models. Mathematical statistics is used to identify types of diseases, calculate prevalence of epidemics, and for other purposes.

[illegible] it is necessary to have information about physical properties of materials from which they are made. Thus in manufacture of prosthetic appliances (false teeth, dentures, valves, etc.) it is essential to know mechanical strength, resistance to multiple loads, elasticity, thermal conductivity, electrical conductivity, and other physical properties.

In some cases it is important to know physical properties of biological systems in order to evaluate their viability or ability to withstand certain external impacts. It is possible to observe some disease manifestations [illegible] changes of physical properties of biological objects.

Physical properties and characteristics of the environment

[illegible]

[illegible]

Medicine and [illegible]

[illegible]

Section 1

MATHEMATICAL TREATMENT OF MEASUREMENT RESULTS. FUNDAMENTALS OF CYBERNETICS

Knowledge of mathematics and physics is also important for a physician because they contribute to formation of a materialistic viewpoint on a living body and the processes in it.

In practice, a physician constantly deals with a number of quantitative indices (body temperature, blood pressure, medication dosage, etc.), so one should know how these values are obtained, what is their accuracy, what units they are presented in. Nowadays, a health professional should have an idea about processing of measurement results, be able to use computers and peripherals. Society, human, computer are examples of systems that are able to receive and process information. Such systems, both living and non-living, are subject of cybernetics studies. Problems of cybernetics as some common problems related to various sections of the course, are discussed here. Thus, this section can be considered as a metrological, mathematical and cybernetics introduction to medical and biological physics.

Chapter 1

Introduction to metrology

Metrology is a science of measurements, methods, and means of ensuring their unity, and ways to achieve the required accuracy. In the chapter, along with general problems of metrology, indicators of measurements in biology and medicine are considered.

§ 1.1. BASIC PROBLEMS AND CONCEPTS OF METROLOGY

A measurement is finding a value of a physical quantity experimentally using technical facilities. Measurements allow us to establish laws of nature and are an element of knowledge of the world around us.

There are direct measurements, when a result is obtained directly by measurement of the value itself (taking a body temperature with a clinical thermometer, measurement of the length of an object with a ruler), and indirect ones, when the sought for value is found by the known relationship between it and values being measured directly (determination of body weight taking into account the buoyant force, determination of the viscosity of liquid by velocity of falling ball in it). Technical facilities of measurements (gauges) can be different. For a reader, the most known gauge is a measurement instrument, where the information being measured is presented in a form that is accessible to direct perception by the observer (the temperature is represented by the thermometer as the length of the mercury column, the current is represented with the ammeter pointer readout or the digital value).

Gauges also include a standard measure that is intended to reproduce the physical quantity of a given size (weight of a certain mass).

One of the most common gauges is a measuring transducer (sensor).

It is designed to generate a signal of the information being measured in a form which is convenient for transmission, further conversion, processing and (or) storage (thus, temperature can be represented by an electrical signal, see § 15.6).

The value of the physical quantity obtained for measurement is different from the true one. The degree of approximation of a measurement findings to the true value of measured parameter is characterized by *accuracy of measurements*. Accuracy of measurement is a quality index of measurements.

Quantitative estimation of a measurement result is given not by its accuracy, but by ***an error***. It is a deviation of the current measurement results from the true value of the quantity being measured. The smaller an error, the higher is measurement accuracy is.

Errors are explained by imperfection of a gauge, inexperience of operator, influence of external factors, etc. Among these causes we can distinguish those that appear irregularly and upon repeated measurements have a different quantitative impact on the result. Such factors generate random errors. These are random quantities, so they can be processed, analyzed and thus taken into account using the appropriate mathematical apparatus: probability theory and mathematical statistics (see Chapters 2 and 3).

Information on the theory of errors required to medical students is given in [1].

One of the basic metrological concepts is units of physical quantities. ***The unit of physical quantity is the conventional physical quantity for quantitative assessment of a corresponding physical quantity.***

Units of physical quantities are mainly grouped into systems of units. The main is the International system of units (SI). Reference material on units of physical quantities is given in [2]. Without dwelling on these questions, let us consider only relative and logarithmic quantities.

In measurement practice, relative values are quite common. They are a relation of one physical quantity to another physical quantity of the same name taken as the initial one. For instance, these are concentration of a solution, relative dielectric permittivity and magnetic permeability, energy conversion efficiency, relative deformation, coefficient of friction, viscosity of the blood relative to viscosity of water, etc.

The relative value does not have any dimension or name. However, in a number of cases, a relative value is traditionally expressed with one hundredfold or thousandfold increase. In this case, it will be presented as a unit, respectively, with percent (%) or per mille (‰).

To express sound pressure level, intensity level of a sound, gain level of an electrical signal, expression of a frequency range, etc. it is more convenient to use the logarithm of a relative value (the most common is a decimal logarithm):

$$L = \lg\frac{a_2}{a_1},$$

where a_1 and a_2 are physical quantities of the same name.

The unit of the logarithmic value is a Bel (B):

$$1\ \mathrm{B} = \lg\frac{a_2}{a_1} \text{ with } a_2 = 10a_1,$$

if a is energy quantity (force, intensity, energy, etc.), or

$$1\ \mathrm{B} = 2\lg\frac{a_2}{a_1} \text{ with } a_2 = \sqrt{10}a_1,$$

if a is force value (force, mechanical stress, pressure, electric field intensity, etc.).

Quite common submultiple unit is a decibel (dB):

$$1\ \mathrm{dB} = 0.1\ \mathrm{B}.$$

§ 1.2. METROLOGICAL ASSURANCE

Measurements are carried out using technical devices. Measurement results should meet certain accuracy and be the same, if identical values are measured, irrespective of whether the measurements are made simultaneously or at different times, in the same laboratory or in different ones.

To meet these requirements the corresponding ***metrological assurance*** should be done. Metrological assurance is establishment and application of scientific and organizational foundations, technical devices, rules and norms, which are required to achieve uniformity and required accuracy of measurements.

The organizational basis of metrological support in our country is the Federal Agency on Technical Regulation and Metrology (Rosstandart), consisting of state and departmental metrological services.

Uniformity of measurements is understood as equality of the results of identical measurements irrespective of place an time of their doing as well as reliability of measurements. The uniformity of measurements make it possible to compare results obtained by different the same type gauges.

To determine errors of gauges and determine their service ability they are ***standardized***. This term ***standardization*** is specific for metrology, although in everyday life it corresponds to the concept of ***checking***. By means of standards and standard gauges, metrological authorities carry out standardization.

A ***standard*** is a gauge or a complex of gauges assuring reproduction and storage of the legalized unit of a physical quantity. Primary standards in our country provide the highest accuracy of reproduction of a unit given. In addition to the primary standards, there are secondary standards, from which the unit is transmitted to the standard gauges. As an example, the standard of light is shown in fig. 27.13.

A standard gauge is a gauge that passed certification (certification is a documentary confirmation that a gauge meets its purpose) as a model one and is used for standardization of working gauges.

Working gauges are gauges used for practical measurements in various fields of life.

Thus, the metrological chain, through which the value of a unit of a physical quantity is transmitted, consists of the following basic links: standards–standard gauges–working gauges.

§ 1.3. MEDICAL METROLOGY. SPECIFIC CHARACTER OF MEDICAL AND BIOLOGICAL MEASUREMENTS

The general term "medical equipment" refers to technical devices used in medicine. Most of medical equipment is medical apparatus, which, in turn, is divided into medical devices and medical appliances.

A ***medical device*** is considered a technical device designed for diagnostic or therapeutic measurements (clinical thermometer, sphygmomanometer, electrocardiograph, etc.).

A ***medical appliance*** is a technical device making it possible to generate energy for therapeutic, surgical or antibacterial energy exposure and assure a composition required of various materials (UHF therapy appliance, electrosurgery, artificial kidney, cochlear prosthetic appliance, etc.) for medical purposes.

Metrological requirements to medical devices as gauges are quite obvious. Many medical devices are intended to have dosing energy deposition on a body, so they are also included in the scope of attention of metrological service.

Measurements in medicine (medical or biomedical ones) as well as the corresponding gauges are quite specific. This peculiarity makes to allocate an individual direction in metrology — *medical metrology*.

Let us consider some problems, which are typical of medical metrology and, to a certain extent, of medical device engineering.

1. Nowadays, most often, medical staff (physician, nurse) that is not technically trained carry out medical measurements. Therefore, it is expedient to create medical devices graduated in units of physical quantities, whose values are the final medical measurement information (direct measurements).

2. Time of measurement is desirable, up to the result, to be as short as possible, and the information to be as complete as possible. These contradictory requirements are met by measuring complexes including computers.

3. It is important to take into account medical indications for metrological graduation of a medical device being created. A physician should determine the sufficient accuracy of results presentation to make any diagnostic conclusion. In this case, one should take into account possible deviations of these indications in individual patients.

4. Many medical devices output information onto a recording medium (e.g. an electrocardiograph), so errors which are characteristic for this or that form of recording should be taken into account (see § 21.5).

5. Another problem is terminological one. According to the requirements of metrology, in the name of the measuring device, physical quantity or unit should be specified (ampere-meter (ammeter), voltmeter, frequency meter, etc.). Names of medical devices do not correspond to this principle (electrocardiograph, phonocardiograph, rheograph, etc.). So an electrocardiograph should be called a millivoltmeter with registration unit (or a recording millivoltmeter).

6. In a number of medical measurements, there can be insufficient information on the relationship between the directly measured physical quantity and the corresponding medical and biological parameters. Thus, in the clinical (bloodless) method of the blood pressure measurement (see § 11.4) it is assumed that the air pressure inside the cuff is approximately equal to the blood pressure in the brachial artery. In fact, this relationship is not so simple and depends on a number of factors, including degree of muscle relaxation. Laboratory measurements (*in vitro*) may differ from the values of the corresponding parameter in the body (*in vivo*).

7. In the process of measurement, medical and biological parameters can change. In practice of applied-physics measurements, they intend to make a few readings to rule out (register) random errors; this is useful in cases where there is confidence in the stability of a physical parameter in measurement. The parameters of a biological system can change over prolonged measurements very much, e.g., due to psychophysiological factors (environmental impact: room conditions, gauge, staff, etc.) or muscle fatigue in case of multiple measurements with a dynamometer. Mobility of organs or an object itself can also generate different measurement results.

Naturally, when developing medical equipment, other requirements (sanitary and hygiene, safety and reliability problems, etc.) should be taken into account; some of them are considered below.

§ 1.4. PHYSICAL MEASUREMENTS IN BIOLOGY AND MEDICINE

Most measurements in medicine are those of physical or physicochemical quantities.

In quantitative diagnostics, these are blood pressure, time dependence of biopotentials, optical power of an eye, etc. In laboratory tests, these are blood viscosity, concentration of sugar in urine, etc. In a process of treatment, it is important to know the dose of ionizing radiation, current for galvanization, the intensity of ultrasound, etc.; absence of such information can not only reduce the medical efficiency but harm too. Quantitative estimation of environmental parameters near a person (air humidity, temperature, and atmospheric pressure) is a necessary condition for prophylaxis and climatic treatment.

All kinds of medical and biophysical measurements can be classified either by their functional basis or by belonging to a corresponding branch of physics. Physical classification is closer to the structure of this course, that is why it is given below.

Mechanical measurements: anthropometric body data, displacement, velocity and acceleration of body parts, blood, air, acoustic measurements, blood and liquid pressure in the body and atmosperic air, vibration measurement, etc.

Thermophysical measurements: temperature of organs, body parts and environment, calorimetric measurements of biological objects, foodstuffs, etc.

Electrical and magnetic measurements: biopotentials, magnetic induction in the human heart, measurement of impedance of biological objects for diagnosing, parameters of electromagnetic fields, and concentration of ions for hygienic purposes, etc.

Optic measurements: colorimetric measurements, measurements of optical characteristics of eye media for diagnostic purposes, spectral measurements for diagnostic and forensic purposes, measurement of characteristics of UV, IR, and visible light for hygienic purposes, etc.

Atomic and nuclear measurements: measurement of ionizing radiation level (dosimetry), etc.

In addition, other physical and chemical measurements can be enumerated: quantitative determination of inhaled and exhaled air composition, gas composition of the blood, pH of the blood and other biological media.

The functional principle of classification of biomedical measurement methods is illustrated by an example of measurement of the cardiovascular system parameters. In this, there are mechanical (ballistocardiography, phonocardiography, measurement of blood pressure), electric and magnetic (electrocardiography, magnetocardiography), optical measurements (oxymetry). It is possible to use other physical methods; e.g., the method of nuclear magnetic resonance (NMR) is used to determine the velocity of blood flow, and many other methods.

Chapter 2

Elements of probability theory

The probability theory studies objective laws related to random events, random variables, and random processes. Physicians rarely think that making diagnosis has a probabilistic character. As it was wittily observed, only a post mortem examination can reliably determine the diagnosis of a person who has died.

§ 2.1. EXPERIMENT WITH MULTIPLE OUTCOMES. RANDOM EVENT

The probability theory studies laws inherent to experiments *with multiple outcomes.* This is a term for experiments, whose results are not possible to *foresee accurately.* E.g., when someone plays roulette, the ball thrown on the rotating wheel can stop in *any* of 37 numbered slots (0, 1, ..., 36), but until the wheel stops the slot number remains unknown.

Experiment and its outcomes

The concepts of "*experiment*" and "*outcome*" are the primary concepts of probability theory.

***An experiment* is a sequence of actions to be done under certain conditions.**

***An outcome* means what is *directly* obtained because of the experiment.**

Experiment is determined if the conditions of the experimentation are specified and the set *of all* its possible *outcomes* is known; the latter is denoted by the letter Ω. For example, for playing *roulette*, the croupier winds the game wheel round, throws the ball on it, waits for the wheel to stop and announces the number of the slot which the ball is located in. The foregoing actions are a description of an *experiment. The experiment outcome* is the announced number of the slot. The set of *all* possible *outcomes* consists of 37 numbers: $\Omega = \{0, 1, 2, ..., 36\}$.

Note that because in each experiment there appears **only one** of all possible outcomes.

In medical research, an experiment is any examination of a patient, e.g., determination of glucose content in his or her blood taken from the vein. The *outcome* is the result of examination.

Random event

Individual outcomes of experiment, as a rule, do not have independent significance. Some of their *sets,* which are called *events,* are of practical interest. E.g., a roulette player can bet his or her money on "even". He *wins* if the ball stops in the slot with an *even* number, and *loses* otherwise. The *specific* number of a slot does not matter. In this case, there are two events of practical interest: "win" is getting an even number, and "loss" is getting an odd number. Nothing else matters.

Outcomes of medical research are also grouped into significant events. E.g., 3 events are considered for determining blood glucose content: this index is *normal* (3.9–6.4 mmol/L), *below normal, above normal.* But the specific value of the index (e.g., 5.18 mmol/L) does not have practical importance. In this example, the event "normal" is the set of all numbers within the interval (3.9–6.4 mmol/L).

***A random event* or simply *an event* is a set of experiment outcomes with a practical interest. Such outcomes are called *conducive* to this event (or *favorable* for it).**

An event occurs if the result of experiment is one of *favorable outcomes.*

In probability theory, capital Latin letters (A, B, C...) denote random events.

§ 2.2. OPERATIONS ON EVENTS. OPPOSITE EVENT. INCOMPATIBLE EVENTS

In order to explain *what this event is,* it is necessary to enumerate all possible outcomes of the experiment (Ω) and designate *favorable* ones. In some cases, it is simple to do, and in other cases, it is much more difficult.

For example, in the experiment the shooter is to fire *one* shot at a target. In this case, only *two* outcomes are possible: A (hit) or B (miss). These outcomes are the simplest events.

Now consider an experiment when the shooter fires *two* shots at the target. In this case, *four elementary* outcomes are possible:

1) A_1 and A_2 — two hits;
2) A_1 and B_2 — a hit and a miss;
3) B_1 and A_2 — a miss and a hit;
4) B_1 and B_2 — two misses.

The event C consisting in the fact that the target is hit by two shots is favored by three outcomes, in which there is at least one hit:

$$C = \{(A_1 \text{ and } A_2), (A_1 \text{ and } B_2), (B_1 \text{ and } A_2)\}.$$

To describe *complex* events, they are presented as a result of operations on simpler events. Such operations are *addition* and *product* of events.

The *sum*, or *union*, of two events A and B is the event that is occurrence of at least one of them.

The *sum* of events is denoted as follows: $A + B$. (In some textbooks, the *sum* of events is denoted as $A \bigcup B$.)

The event $A + B$ is the set of outcomes which are favorable to *at least one* of events A, B.

***The product*, or *intersection*, of two events A and B is called an event consisting in occurrence of both events.**

The *product* of events is denoted as follows: $A \cdot B$. (In some textbooks *intersection* of events is denoted as $A \bigcap B$.)

The event $A \cdot B$ represents a set of outcomes, favorable *for each* event (both for event A and for event B).

The foregoing complex event C, which is the hit on the target by two shots, is written as operations of addition and multiplication of simple events (A is the hit, B is the miss) in the following way:

$$C = A_1 \cdot A_2 + A_1 \cdot B_2 + B_1 \cdot A_2.$$

Let us analyze a simple example that explains the technique of performing operations of event addition and multiplication. A dice is thrown. Event A is even number falling: $A = \{2, 4, 6\}$. Event B is falling of a number that is a multiple of three: $B = \{3, 6\}$.

- **Addition:** $A + B$ is a number that is **either** even **or** is divided by 3: $A + B = \{2, 3, 4, 6\}$.
- **Product:** $A \cdot B$ is a number that is **both** even, **and** is divided by 3: $A \cdot B = \{6\}$.

It is convenient to illustrate operations on events graphically with special Venn diagrams. In them, the space of elementary outcomes Ω is designated with a circumference, whose points are interpreted as elementary outcomes. Simple events are designated by some figures, e.g., ovals. The image of the sum and the product of events is shown in fig. 2.1 (the dark area).

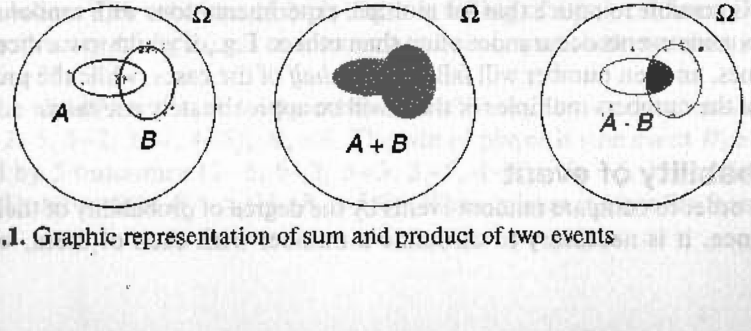


Fig. 2.1. Graphic representation of sum and product of two events

Opposite event

To each event A, it is possible to map the opposite event $\overline{A}$ (is read "not A"), consisting of all outcomes, *unfavorable* for A. A graphic illustration of events A and $\overline{A}$ is represented in fig. 2.2.

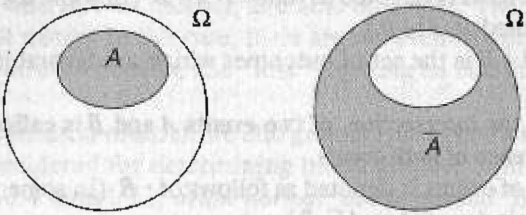


Fig. 2.2. Event A and event $\overline{A}$ which is opposite to the former

The event, which is *opposite* to event A, is the following: for experimentation, event *A did not occurr.*

Let us note that $A + \overline{A} = \Omega$.

Incompatible events

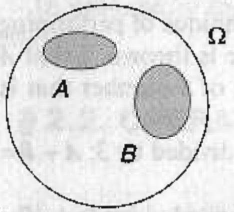


Fig. 2.3. Incompatible events do not have common outcomes

Incompatible events A are of great importance in the probability theory.

***Incompatible events* are events that cannot happen simultaneously (for one experimentation).**

Incompatible events *do not have a common outcome* that is why they are represented by non-intersecting figures (fig. 2.3).

An important special case of incompatible events is initial and opposite events (A and $\overline{A}$).

§ 2.3. CLASSIC DEFINITION OF PROBABILITY. AXIOMS OF PROBABILITY THEORY

It is possible to notice that for multiple experimentations with random outcomes some events occur more often than others. E.g., if you throw a dice many times, an even number will fall in about *half* of the cases, while the proportion of the numbers multiples of three will be approximately *one third.*

Probability of event

In order to compare random events by the degree of probability of their occurrence, it is necessary to associate a number with each of them, which is the greater the more possible this event is. This number determines the *probability* of the event.

***Probability of an event* is a quantitative characteristic of the possibility of its occurrence.**

Probability is denoted by the letter "P": the probability of event A is denoted by $P(A)$ or P_A.

The theory of probability was initially invented for analysis of games of chance and was applied to experiments, whose all outcomes are *equally possible*.

Outcomes of experiment are called *equally possible*, if no objective reasons by virtue of some outcomes can be more given than others.

E.g., due to the symmetry of a dice, the chances of all its faces falling are *equal*. Therefore, a throw of a dice is an experiment with equally possible outcomes.

Classical definition of probability

Let us consider an experiment with N equally possible outcomes. Let us denote the number of outcomes, which are favorable for event A, as N_A.

***The probability of a random event* is the ratio of the number of favorable outcomes for this event to the number of all equally possible outcomes of this experiment:**

$$P_A = N_A/N. \tag{2.1}$$

Historically, this formula was given the name "Classical Definition of Probability". It was the first quantitative result of a proposed theory which made it possible to determine probabilities of success in various kinds of games of chance. Let us consider the application of this definition to dice game.

Problem. Players A and B play by throwing two dices each. Player A wins when the sum of his or her points is seven. Player B wins when the sum of his or her points is eight. Who benefits from this game?

Solution. The outcome of each throw is falling of a *pair* of faces. Due to the symmetry of dices, all outcomes are equal, and their number is $N = 6 \cdot 6 = 36$.

The win of the player A (the event A) 6 is favored by six outcomes (1–6, 6–1, 2–5, 5–2, 3–4, 4–3); $N_A = 6$. The win of player B (the event B) 6 is favored by 5 outcomes (2–6, 6–2, 5–3, 3–5, 4–4); $N_B = 5$. Using formula (2.1), let us find $P_A = 6/36$, $P_B = 5/36$. Thus, player A benefits from this game.

Axioms of the probability theory

Not all experiments have equally possible outcomes. For instance, in shooting at a target, the possibilities of hit and miss are obviously different. In order to generalize the concept of probability to arbitrary experiments with random outcomes, it was necessary to introduce a number of general concepts and properties.

The limits within which probability of an event are changes established according to two special concepts.

1. A *certain event* is an event that is *sure* to occur because of an experiment. Such an event is the set of *all* possible outcomes Ω.

2. An *impossible event* is an event that in this experiment cannot occur at all. For instance, in playing roulette, number 38 cannot fall; it is simply not on the wheel. An impossible event is denoted with the symbol $\varnothing$.

The probability of a certain event is taken as *one*:

$$P_\Omega = 1.$$

The probability of an impossible event is taken as *zero*:

$$P(\varnothing) = 0.$$

Two more axioms are added to these properties of probability:

- the probability of any event A lies between zero and one:

$$0 \le P_A \le 1;$$

- the probability of the sum of *incompatible* events is equal to the sum of their probabilities:

$$P(A + B) = P_A + P_B. \qquad (2.2)$$

It can be proved that the probability of the sum of *joint* events is given by the following formula:

$$P(A + B) = P_A + P_B - P(A \cdot B). \qquad (2.3)$$

§ 2.4. RELATIVE FREQUENCY OF AN EVENT, THE LAW OF LARGE NUMBERS

Conditions in which it is permissible to use the classical definition of probability are extremely rare since experiments with *equally possible* outcomes are rather an exception than a rule. If the outcomes *are not equally possible*, then the probability of an event cannot be calculated by formula (2.1).

Let us consider a method of *experimental* evaluation of some event probability A. Let us reiterate the same experiment several times and count in how many experiments this event has *occurred*.

625023, ень, ул. Одесская, д. 51
ФГБОУ ВО ский ГМУ Минздрава России
БИБЛИОТЕКА

A relative frequency of a certain event *A* in a series of accomplished experiments is the ratio of the number of experiments (n_A), in which the event occurred, to the total number of accomplished experiments (n):

$$P_A^* = \frac{n_A}{n}. \tag{2.4}$$

If n is small, the relative frequency of an event is random at a certain extent. However, as the number of experiments increases, the frequency tends to *stabilize*, approaching, with negligible fluctuations, a certain constant. The table below shows how the frequencies (P^*) of falling of tails change upon an increase in the number of throws (n) for a symmetrical coin.

Table 2.1

n	10	50	75	100	200	300	400	500	600
P^*	0.400	0.540	0.493	0.510	0.505	0.503	0.498	0.502	0.499

The plot curve corresponding to these changes is shown in fig. 2.4.

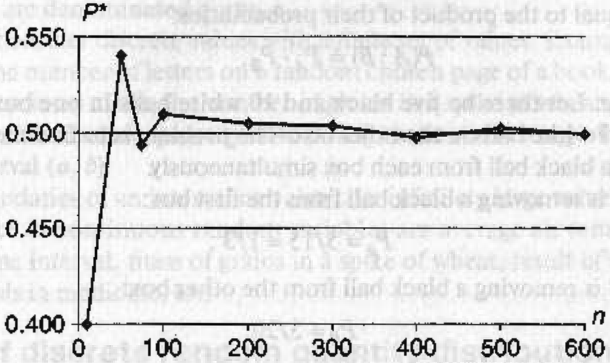


Fig. 2.4. Convergence of the relative frequency of the event to its probability

The relative frequency of an event and its probability are connected by the ***law of large numbers.***

As the number of experiments increases unlimitedly, the frequency of the event tends to its probability:

$$\frac{n_A}{n} \to P(A) \text{ with } n \to \infty. \tag{2.5}$$

This ratio is sometimes called the statistical probability definition. In accordance with the law of large numbers, the probability of an event can be taken as its relative frequency with a large number of experiments.

§ 2.5. INDEPENDENT EVENTS. ADDITION AND MULTIPLICATION OF PROBABILITIES OF INDEPENDENT EVENTS

The concept of *statistical independence* occupies an important place in the probability theory and is defined as follows:

Events *A* and *B* are called ***independent***, if the fact of occurrence of one of them does not change the probability of occurrence of the other.

A typical example of *independent* events is events that occur in experiments with *independent outcomes.*

Two experiments are called ***independent* if the outcome of one experiment cannot influence the outcome of the other.**

E.g., if you throw two dices, the result of the first throw does not affect the result of the second throw.

For independent events, the *theorem of multiplication of probabilities* is applicable.

The probability of an event, that is the product of independent events *A* and *B*, is equal to the product of their probabilities:

$$P(A \cdot B) = P_A \cdot P_B. \tag{2.6}$$

Example. Let there be five black and 10 white balls in one box, and three black and 17 white balls in the other box. The problem is to find the probability of drawing a black ball from each box simultaneously.

Event *A* is removing a black ball from the first box:

$$P_A = 5/15 = 1/3\ .$$

Event *B* is removing a black ball from the other box:

$$P_B = 3/20.$$

Event $A \cdot B$ is both balls being black:

$$P(A \cdot B) = P_A \cdot P_B = 1/3 \cdot 3/20 = 1/20.$$

Application of the probability multiplication theorem to formula (2.3) implies the following law of finding of two independent events sum probability:

$$P(A + B) = P_A + P_B - P_A \cdot P_B. \tag{2.7}$$

Example. Let there be five black and 10 white balls in one box, and three black and 17 white balls in the other box. The problem is to find the probability of drawing *at least one* black ball upon removing a ball from each box. Using values P_A, P_B and $P(A \cdot B)$, obtained in previous example, we find:

$$P(A + B) = 1/3 + 3/20 - 1/20 = 22/60.$$

§ 2.6. DISCRETE AND CONTINUOUS RANDOM QUANTITIES. DISTRIBUTION SERIES, DISTRIBUTION FUNCTION. PROBABILITY DENSITY

Often, numerical values are connected with outcomes of some experiment. E.g., numbers are on the faces of a cube, so falling of any face is falling of the corresponding number. When you throw the same cube again, the numbers will change randomly. In this case, we speak about a random quantity.

Under a *random quantity* (RQ) we mean a quantity the amount of which depends on results of an experiment with random outcomes.

Random quantities are denoted with capital letters (X, Y...), and their values with lowercase letters (x, y...).

Of the multitude of all random quantities, two most common types are distinguished: *discrete* and *continuous ones*.

***Discrete* random quantity is such RQ that can take only a finite (or countable) set of values.**

These values are numbered x_1, x_2, x_3... and the probabilities of their appearance are denominated p_1, p_2, p_3...

We will consider discrete values with a *finite* set of values. Examples of such values are the number of letters on a random chosen page of a book, the energy of an electron in an atom, the number of grains in a spike of wheat, etc.

A *continuous* random quantity is such RQ that can take any value in some specific interval (a, b).

The boundaries of an interval can also take infinitely large values.

Examples of continuous random variables are average air temperature in a certain time interval, mass of grains in a spike of wheat, result of any quantitative analysis in medicine, etc.

Series of discrete random quantity distribution

A discrete random quantity is considered preassigned if all its possible values x_1, x_2...x_N and their corresponding probabilities p_1, p_2...p_N are known. The set of RQ values and their probabilities, specified in the form of a table, is called *distribution series*, or *distribution* of a discrete random quantity:

X	x_1	x_2	x_3	...	x_N
P	p_1	p_2	p_3	...	p_N

The sum of all probabilities is one:

$$\sum_{i=3}^{N} P_i. \tag{2.8}$$

A distribution series is the most complete characteristic of a *discrete* RQ.

Distribution function

The complete characteristic of a continuous random quantity is the *distribution function* $F(x)$, the value of which in each point x is equal to the probability of the random quantity X to take a value less than x:

$$F(x) = P(X < x). \tag{2.9}$$

The probability that the RQ value is less than1 is 0 (the fact of all numbers which are less than $+\infty$ is a certain event), so $F(+\infty) = 1$. The probability of the fact that a value of RQ will be less than $-\infty$, is zero (the fact that there are no such numbers means an impossible event), that is why $F(-\infty) = 0$. A typical form of a distribution function is shown in fig. 2.5.

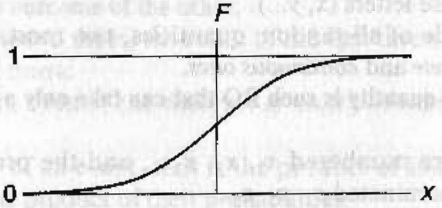


Fig. 2.5. Typical distribution function of a random quantity

The distribution function makes it possible to calculate the probability of a value of a continuous random quantity to fall within the given interval (x_1, x_2):

$$P(x_1 < X < x_2) = F(x_2) - F(x_1). \tag{2.10}$$

Distribution density

Distribution functions of all continuous random quantities are very much alike: they all increase uniformly from 0 to 1. Individual specialties of random quantities are revealed by another function called *distribution density*.

Distribution density **(or probability density)** $f(x)$ **of a continuous random quantity is the derivative of the initial distribution function:**

$$f(x) = \mathrm{d}F/\mathrm{d}x. \tag{2.11}$$

The distribution density has the following probabilistic interpretation:

The probability the continuous random quantity X **to take values within a small interval** $(x, x + dx)$**, is equal to the product of the probability density by the width of the interval:**

$$\mathrm{d}P = f(x) \cdot \mathrm{d}x. \tag{2.12}$$

For plotting a density distribution, the probability of an experiment the value of a continuous random quantity falling within a given interval (x_1, x_2), is equal to the area of the corresponding curvilinear trapezoid (fig. 2.6). For all that, the area under the entire plot curve is equal to *one*. This condition is equivalent to the normalization condition (2.8) for discrete RQs.

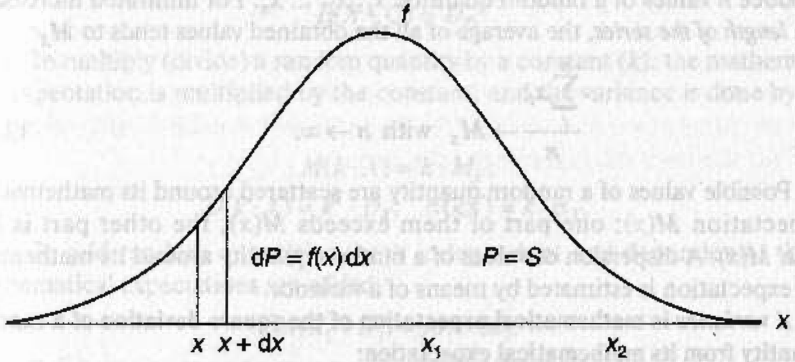


Fig. 2.6. Typical distribution density of a random quantity

For problems of practical statistics, only three types of intervals are of interest: the "left tail" of distribution $(-\infty, x_1)$; "central" interval (x_1, x_2) and "right tail" of distribution $(x_2, +\infty)$.

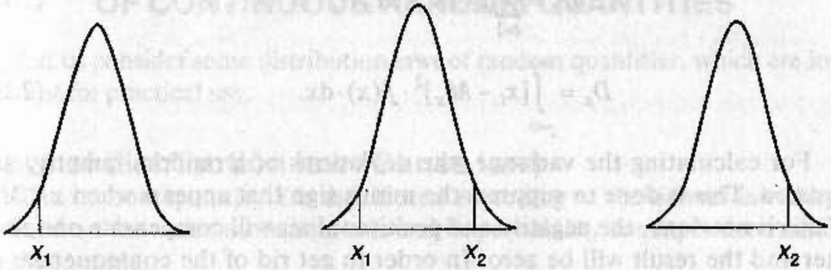


Fig. 2.7. Intervals used in practical statistics

§ 2.7. NUMERICAL CHARACTERISTICS OF RANDOM QUANTITIES

The distribution series and the distribution density contain full information about the corresponding random quantity; nevertheless, in solving many practical problems, it is enough to know two numerical characteristics of the random quantity: *mathematical expectation* and *dispersion*. We will give a not very strict but clear definition of these characteristics.

***The mathematical expectation* M_X of a random quantity X is its arithmetic mean.**

This definition has the following meaning. Let a series of n experiments produce n values of a random quantity: $x_1, x_2, \ldots x_n$. For unlimited increase in the *length of the series*, the average of all the obtained values tends to M_X:

$$\frac{\sum_{i}^{n} x_i}{n} \to M_x \text{ with } n \to \infty. \tag{2.13}$$

Possible values of a random quantity are scattered around its mathematical expectation $M(x)$: one part of them exceeds $M(x)$, the other part is less than $M(x)$. A dispersion of values of a random quantity around its mathematical expectation is estimated by means of a variance.

***A variance* is mathematical expectation of the square deviation of a random quantity from its mathematical expectation:**

$$D_x = M[X - M_X]^2. \tag{2.14}$$

The formulas for calculating the variance of discrete and continuous random quantities are as follows:

$$D_x = \sum_{i=1}^{n} p_i \cdot [x_i - M_x]^2, \tag{2.15}$$

$$D_x = \int_{-\infty}^{+\infty} [x_i - M_x]^2 \cdot f(x) \cdot dx. \tag{2.16}$$

For calculating the variance, the deviations of a random quantity are squared. This is done to suppress the minus sign that appears when $x < M_X$. If this is not done, the negative and positive values will compensate one another and the result will be zero. In order to get rid of the consequences of squaring deviations, after calculating the variance, a square root is extracted from it. The resulting value is used as a measure of deviation of a random quantity from the mean value.

***The quadratic deviation* (QD) of a random quantity is the square root of its variance:**

$$\sigma_x = \sqrt{D_x} \tag{2.17}$$

(sometimes the term "standard deviation" is used).

For data processing, mathematical operations are done on random quantities, because of which new random quantities are obtained. Let us show how mathematical expectations and variances change in this case.

1. To add a random quantity with a constant (C), the latter is added to the mathematical expectation and the variance, and QD do not change:

$$M(X+C)=M_X+C;$$
$$D(X+C)=D_X.$$

2. To multiply (divide) a random quantity by a constant (k), the mathematical expectation is multiplied by the constant, and the variance is done by its square:

$$M(k\cdot X)=k\cdot M_X;$$
$$D(k\cdot X)=k^2\cdot D_X,\quad \sigma(kX)=k\cdot\sigma_X.$$

3. To add random quantities (both independent and dependent), their mathematical expectations are added:

$$M(X_1+X_2)=M_1+M_2.$$

4. To add the *independent* random quantities, their variances are added:

$$D(X_1+X_2)=D_1+DX_2.$$

§ 2.8. SOME DISTRIBUTION LAWS OF CONTINUOUS RANDOM QUANTITIES

Let us consider some distribution laws of random quantities, which are important for practical use.

Normal distribution law (Gauss law)

The random quantity X is distributed according to the *normal* law if it is defined on the entire numerical axis and its probability density is determined by the formula:

$$f(x)=\frac{1}{\sigma_X\sqrt{2\pi}}\cdot\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right); \qquad (2.18)$$

where $\mu=M_X$ is mathematical expectation of the random quantity; σ is its quadratic deviation.

For practical statistics, importance of the normal distribution law is related to the *Central Limit Theorem* according to which the sum of a large number of independent random quantities with the same distribution law has a distribution that can be considered normal. At the same time, the law of distribution to which summands are subject does not matter and can be totally unknown. We will use this property in the next paragraph.

Fig. 2.8 shows plot curves of the probability density of two normally distributed RQ with $\mu = 0$, $\sigma = 2$ and $\mu = 2$, $\sigma = 1$.

Let us note some properties of these plot curves:

- the plot curve of density distribution of the normal law is symmetric and bell-shaped; the line of symmetry passes through the mathematical expectation point of a random quantity ($x = \mu$);
- in the point $x = \mu$ the function attains its maximum value;
- the parameter σ characterizes the shape of the distribution curve: the less σ, the narrower and higher is the plot curve.

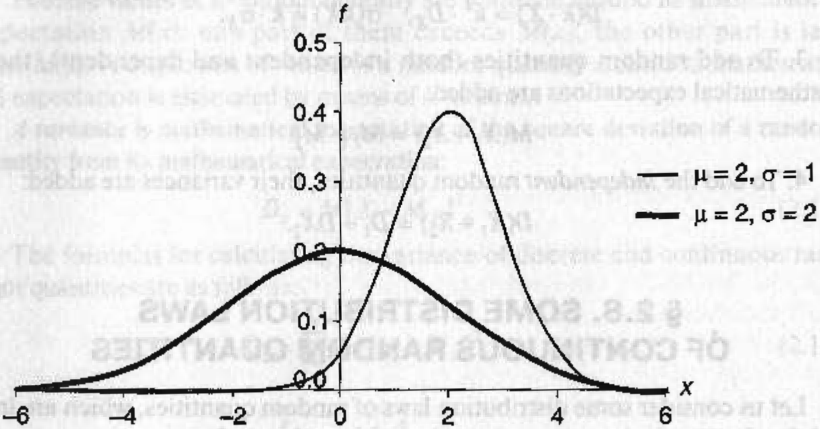


Fig. 2.8. Probability density plot curves for the normal distribution law

Standard computer functions are used to calculate values of the distribution function and the probability density of the normal law values. In the well-known Excel application, these calculations are done by the statistical function NORMDIST (x, μ, σ, m). When $m = 0$, the calculation is done for the *distribution density*, and for $m = 1$ the calculation is done for the *distribution function*.

The normal distribution with $\mu = 0$ and $\sigma = 1$ is called *standard*.

Using the properties of mathematical expectation and variance it is not difficult to show that if random quantity X does not have normal distribution with parameters μ and σ, then random quantity $X_0 = (X - \mu)/\sigma$ has the *standard normal* distribution. Hence, the probability of event $|X - \mu| < k \cdot \sigma$ is equal to the probability of event $|X_0| < k$. Using formula (2.10), we will find

$$P(-k\sigma < |X - \mu| < k\sigma) =$$
$$= \mathrm{NORMDIST}(k, 0, 1, 1) - \mathrm{NORMDIST}(-k, 0, 1, 1).$$

For $k = 1$, $k = 2$ and $k = 3$ we obtain:

$$\begin{aligned} P(-\sigma < X - a < \sigma) &= 0.6826, \\ P(-2\sigma < X - a < 2\sigma) &= 0.9544, \\ P(-3\sigma < X - a < 3\sigma) &= 0.9974. \end{aligned} \qquad (2.19)$$

The last number shows that the probability of deviation of a normally distributed random quantity from the average by more than 3σ amounts to only 0.26%. Relations (2.19) are shown in fig. 2.9.

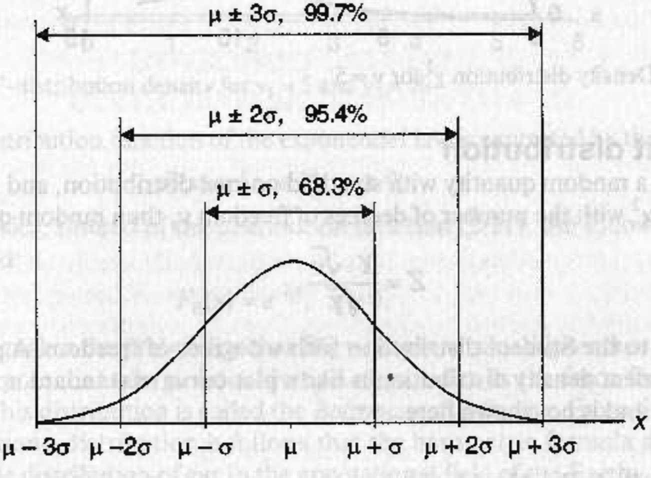


Fig. 2.9. Probabilities of deviation of a normally distributed random quantity from the mathematical expectation

Distribution χ^2, Student distribution and Fisher distribution

The three distributions below are connected with the standard normal distribution, which are of great importance in mathematical statistics.

Distribution χ^2

Let $X_1, X_2, \dots X_\nu$ be independent random quantities with a *standard normal distribution*. Then the sum of their squares is subject to distribution χ^2 (chi-square):

$$Y = X_1^2 + X_2^2 + \dots + X_\nu^2. \qquad (2.20)$$

The number of summands ν (nu) is called the number of degrees of freedom. Density distribution plot curve for $\nu = 5$ is shown in fig. 2.10.

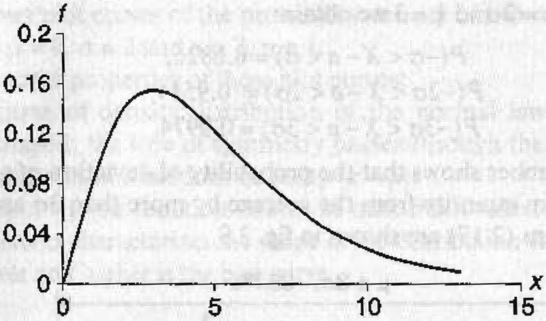


Fig. 2.10. Density distribution χ^2 for $\nu = 5$

Student distribution

If X is a random quantity with standard normal distribution, and Y has distribution χ^2 with the number of degrees of freedom ν, then random quantity

$$Z = \frac{X \cdot \sqrt{\nu}}{\sqrt{Y}} \tag{2.21}$$

is subject to the Student distribution with ν degrees of freedom. A plot curve of the Student density distribution is like a plot curve of standard normal distribution, and is not shown here.

Fisher distribution

If Y_1 and Y_2 are independent random quantities having χ^2-distribution with ν_1 and ν_2 degrees of freedom respectively, then the relation

$$F = \frac{Y_1 \cdot \nu_2}{Y_2 \cdot \nu_1} \tag{2.22}$$

has F-distribution of Fisher. At all that, ν_1 is called the numerator degrees of freedom, and ν_2 is called the denominator degrees of freedom.

Density distribution plot curve of F-distribution for $\nu_1 = 5$ and $\nu_2 = 10$ is presented in fig. 2.11.

Exponential distribution law. Boltzmann distribution

Continuous random quantity with positive values, whose probability density is given by the formula:

$$f(x) = 1 - \lambda \cdot e^{-\lambda x}, \quad x \geq 0, \tag{2.23}$$

is called distributed according to the *exponential* law.

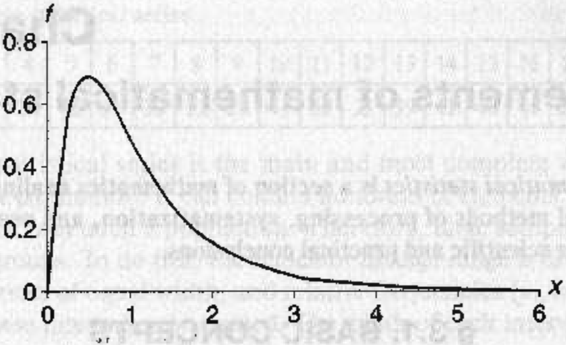


Fig. 2.11. F-distribution density for $\nu_1 = 5$ and $\nu_2 = 10$

The distribution function of the exponential law is expressed by the formula:

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0. \tag{2.24}$$

In physics, instead of the distribution function (2.21), the following function is used:

$$F_E(x) = e^{-\lambda x}; \quad x \geq 0, \tag{2.25}$$

which is equal to the probability that RQ will take a value *higher than* x. By this function, potential energies distribution of particles in the force fields is described. This distribution is called the *Boltzmann distribution.* From the statistical Boltzmann distribution it follows that the barometric formula determing the altitude distribution of gas in the gravitational field of the Earth:

$$n = n_0 \cdot \exp(-mgh/kT), \tag{2.26}$$

where n and n_0 are molecule concentration at altitude h and near the Earth surface; m is the mass of the molecule; k is the Boltzmann constant; T is the absolute temperature.

Chapter 3

Elements of mathematical statistics

***Mathematical statistics* is a section of mathematics dealing with mathematical methods of processing, systematization, and use of statistical data for scientific and practical conclusions.**

§ 3.1. BASIC CONCEPTS OF MATHEMATICAL STATISTICS

In biomedical problems, it is often necessary to study the distribution of this or that indicator for a very large number of individuals. In different individuals, this indicator has a different meaning, so it is a random quantity. For instance, any medical preparation has different efficiency as it is applied to different patients. However, to form an idea about the efficiency of this preparation, no need to apply it to *all* patients. One can monitor results of application of the preparation to a quite small group of patients, and on the base of the data obtained reveal the essential characteristics (efficiency, contra-indications) for the treatment process.

The *general population* is a set of homogeneous elements characterized by some indicator to be studied. This indicator is a *continuous* random quantity with density distribution $f(x)$.

If we are interested in the prevalence of any disease in a certain region, the general population is the entire population of the region. If we want to find out the susceptibility of men and women to this disease separately, we should analyze two general populations.

To study properties of the general population, a certain part of its elements is sampled.

A sample is a part of the general population chosen for examination (treatment).

If it does not cause misunderstanding, a sample is called both a *set of objects* selected for examination, and the *set of values* of studied indicator obtained in the study. These values can be represented in several ways.

A simple statistical series is values of the studied indicator, recorded in the order in which they have been obtained.

An example of a simple statistical series obtained for measuring a surface wave velocity (m/s) in the forehead skin of 20 patients is presented in Table 3.1.

Table 3.1. Simple statistical series

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X	35	32	38	34	42	36	39	33	28	47	39	33	39	41	40	46	28	37	36	38

A simple statistical series is the main and most complete way to record the results of examination. It can contain hundreds of elements. It is very difficult to glance over such a population. Therefore, large samples are usually divided into groups. To do this, the indicator change range is divided into several (N) *intervals* of equal width, and relative frequencies (n_i/n) of indicator falling into these intervals are counted. The width of each interval is equal to:

$$d = (x_{\max} - x_{\min})/N.$$

The interval boundaries have the following values:

$$x_{\min}; x_{\min} + d; x_{\min} + 2d \ldots x_{\max}.$$

If a sample element has the unique value and does not belong to any interval, it is included into the *left* interval. The data grouped together in this way are called *an interval statistical series.*

An interval statistical series is a table showing intervals of indicator values and relative frequencies of indicators found in these intervals.

In our case, it is possible to form, such interval statistical series ($N = 5$, $d = 4$), Table 3.2.

Table 3.2. Interval statistical series

$x, x + \Delta x$	28–32	32–36	36–40	40–44	44–48
$P^* = n_i/n$	0.10	0.25	0.40	0.15	0.10

Here, two values equal to 28 are included in the interval 28–32 (Table 3.1), and values 32, 33, 34, and 35 are included in the interval 32–36.

An interval statistical series can have a graphic representation too. To create it, the intervals of the values of the indicator are put along the abscissa axis,, and on each of them, as on a basis, a rectangle with a height equal to the relative frequency is built. The resulting bar chart is called a *histogram.*

In the histogram, statistical laws of the indicator distribution are seen clearly enough.

In case of a large sample size (several thousand) and a small bar width, the shape of the histogram is close to the shape of the indicator *density distribution* plot curve.

The number of histogram bars can be calculated by the following formula:

$$N = 1 + 3{,}3 \cdot \lg n. \tag{3.1}$$

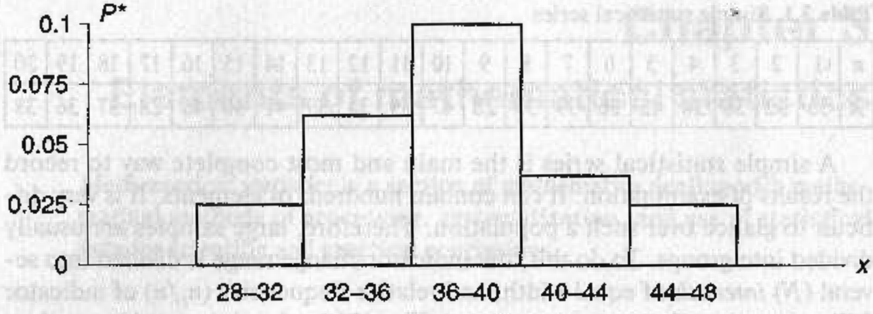


Fig. 3.1. Histogram

Making a histogram manually is a long process. Therefore, for their automatic plotting, there are computer programs.

§ 3.2. NUMERICAL CHARACTERISTICS OF A STATISTICAL SERIES

Many statistical procedures use sample estimates for mathematical expectation and variance (or QD) of the general population.

Sample average ($\overline{X}$) is the arithmetic mean of all elements of a simple statistical series:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}. \qquad (3.2)$$

In our example, $\overline{X} = 37.05$ (m/s).

The sample average is *the best* estimate of the general average *M*.

Sample variance s^2 is equal to the sum of element deviation squares from the sample average divided by $n - 1$:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n-1}. \qquad (3.3)$$

In our example, $s^2 = 25.2$ (m/s)^2.

Note that for calculating the sample variance, the denominator of the formula is not the sample size n, but $n - 1$. This is due to the fact that for calculating deviations in formula (3.3), its estimate, *a sample average*, is used instead of an unknown mathematical expectation.

The sample variance is *the best* estimate of the general variance (σ^2).

Sample quadratic deviation (s) is the square root of the sample variance:

$$s = \sqrt{s^2}. \tag{3.4}$$

In our example, $s = 5.02$ (m/s).

The sample *quadratic* deviation is the best estimate of the general QD (σ).

For unlimited increase in the sample size, all sample characteristics tend to the corresponding characteristics of the general population.

$$\overline{X} \to M,\ s^2 \to \sigma^2,\ s \to \sigma \text{ for } n \to \infty.$$

Computer formulas are used to calculate sample characteristics. In Excel, these calculations are operated by statistical functions AVERAGE, DISP, and STDEVA.

§ 3.3. INTERVAL ESTIMATE

All sample characteristics are *random quantities*. This means that for another sample of the same volume, values of the sample characteristics will be different. Thus, sample characteristics are just *estimates* of the corresponding characteristics of the general population.

The disadvantages of sample estimation are compensated by *interval estimate* resulting in *a numerical interval* within which the true value of the estimated parameter is found with the given probability P_c.

Let U_G be a certain parameter of the general population (general average, general variance, etc.).

An interval estimate of parameter U_G is called an interval (U_1, U_2), meeting the condition:

$$P(U_1 < U_G < U_2) = P_c. \tag{3.5}$$

Probability P_c is called the *confidence probability*.

Confidence probability P_c is the probability that the true value of a quantity being estimated is *within* the indicated interval.

For all that, the interval (U_1, U_2) is called a *confidence interval* for the estimated parameter.

In many cases, instead of a confidence probability, a value connected with it is used: $\alpha = 1 - P_c$, which is called a *significance level*.

A *significance level* is the probability of the fact that the true value of the estimated parameter is *beyond* the confidence interval given.

Sometimes α and P_c are expressed in percent form: 5% instead of 0.05 and 95% instead of 0.95.

For interval estimation, first, the corresponding *confidence probability* is chosen (usually these are 0.95 or 0.99), and then the corresponding value interval of the parameter being estimated is found.

Let us note some general properties of interval estimates.

1. The lower the significance level (the higher P_c), the wider the interval estimate. So, when the significance level is 0.05, the interval estimation of the general average is $34.7 < M < 39.4$, but for the level of 0.01 it will be much wider: $33.85 < M < 40.25$.

2. The greater the sample size n, the narrower the interval estimate with a significance level chosen. Let there be a 5 percent estimate of the general average ($\beta = 0.05$), obtained from a sample of 20 elements, then $34.7 < M < 39.4$.

Having increased the sample size to 80, we obtain a more accurate estimate with the same significance level: $35.5 < M < 38.6$.

In general, creation of reliable confidence estimates requires knowledge of the law, according to which a random indicator being estimated is distributed in the general population. Let us consider how it can create interval estimate of the *general average* of indicator, which is distributed in the general population according to the *normal* law.

§ 3.4. INTERVAL ESTIMATE OF THE GENERAL AVERAGE FOR THE NORMAL DISTRIBUTION LAW

Determination of the interval estimate of the general average M for the general population with the normal distribution law is based on the following property: for a sample of size n the relation

$$t = \frac{(M - X) \cdot \sqrt{n}}{s}$$

is subject to Student distribution with the degrees of freedom $\nu = n - 1$.

Here $\overline{X}$ is a sample average, and s is a sample QD.

Using Student distribution tables or their computer form it is possible to find such a boundary value t_b that with a confidence probability given, the following inequality is correct:

$$-t_b < t < t_b.$$

The following inequality for M corresponds to this one:

$$\overline{X} - \varepsilon < M < \overline{X} + \varepsilon, \quad \varepsilon = \frac{t \cdot s}{\sqrt{n}},$$

where ε is half-width of the confidence interval.

Thus, the confidence interval for M is done in the following sequence:

1. Confidence probability P_c is chosen (usually 0.95 or 0.99), and for it, using Student's distribution table, parameter t_b is found.
2. The half-width of confidence interval ε is calculated:

$$\varepsilon = \frac{t_b \cdot s}{\sqrt{n}}. \tag{3.6}$$

3. The interval estimator of a general average with confidence probability chosen is obtained:

$$P(\overline{X} - \varepsilon < M < \overline{X} + \varepsilon) = P_c. \tag{3.7}$$

Briefly, it is written in the following way:

$$M = \overline{X} \pm \varepsilon, \quad P_c = \ldots \tag{3.8}$$

Computer procedures have been developed to find interval estimates.

Let us explain how to use Student's distribution table. This table has two "inputs": the left column, called the number of degrees of freedom, $\nu = n - 1$, and the top line that is the significance level α. On the intersection of the corresponding row and column, we find Student's coefficient t.

Let us apply this method to our sample. A fragment of Student's distribution table is below.

Table 3.3. A fragment of Student's distribution table

ν	Significance level α						
	0.5	0.4	0.3	0.2	0.1	0.05	0.01
1	1.00	1.38	2.00	3.08	6.31	12.71	63.66
10	0.70	0.88	1.10	1.37	1.81	2.23	3.17
19	0.69	0.86	1.09	1.33	1.73	2.09	2.86

The simple statistical series for a sample of 20 people ($n = 20$, $\nu = 19$) is shown in Table 3.1. For this series, calculations by the formulas (3.1)–(3.3) result: $\overline{X} = 37.05$; $s = 5.02$.

Let us choose $\alpha = 0.05$ ($P_c = 0.95$). On the intersection of line 19 and column 0.05 we find $t = 2.09$.

Let us calculate the accuracy of the estimate by formula (3.6): $\varepsilon = 2.09 \times$ $\times\, 5.02/\sqrt{20} = 2.34$.

Let us create the interval estimate: With the probability of 95%, the unknown general average meets the following inequality:

$$37.05 - 2.34 < M < 37.05 + 2.34, \text{ or } M = 37.05 \pm 2.34 \text{ (m/s)}, P_c = 0.95.$$

§ 3.5. METHODS FOR STATISTICAL HYPOTHESES TESTING

Statistical hypotheses

Before formulating what a statistical hypothesis is, let us consider the following example.

For comparison of two treatment options for some disease, two groups of patients 20 persons each were selected. They received therapy with these methods. For each patient, the *number of manipulations* after which a positive effect was achieved was recorded. Accordingly to these data, for each group sample average ($\overline{X}$), sample variance (s^2) and sample QD (s) were found.

The results are represented in Table 3.4.

Table 3.4

Manipulations number	3	4	5	6	7	8	9	10	11	12	$\overline{X}$	s^2	s
Group 1	1	2	3	7	4	2	1				6.05	2.16	1.45
Group 2	1		2	3	6	3	2	1	1	1	7.45	4.05	2.01

The number of manipulations required to obtain a positive effect is a random quantity, all information about which is currently contained in the sample.

In Table 3.4, we see that the sample average in the first group is less than in the other. Does this mean that there is the same ratio for general averages: $M_1 < M_2$? Are statistical data sufficient for such a conclusion? Answers to these questions are obtained by *statistical hypotheses testing*.

Statistical hypothesis *is an assumption concerning properties of the general population.*

We will consider hypotheses about properties of ***two*** general populations.

If general populations have *known, equal* distributions of the value being estimated, and our assumptions concern values of *some parameter* of this distribution, such hypotheses are called *parametric*. For instance, samples are extracted from general populations with *normal distribution law* and the same variance. It is required to find out whether the general averages of these populations are *the same*.

If the laws of distribution of the general populations are not known at all, then hypotheses about their properties are called ***nonparametric***. For instance, whether the distribution laws of the general populations from which the samples are extracted *are the same*.

Null and alternative hypotheses. Problem of hypotheses testing. Significance level

Let us go into the terminology used for hypotheses testing.

- H_0 is a null hypothesis (hypothesis of a skeptic), the hypothesis *of absence of differences* between the samples being compared. A skeptic believes that the differences between the sample estimates obtained from the research results are random;
- H_1 is an alternative hypothesis (hypothesis of an optimist) which means a hypothesis that there are differences between the compared samples. An optimist believes that the differences between the sample estimates are caused by objective conditions and correspond to the differences of the general populations.

Hypothesis	**Opinion of the hypothesizer**
Null — H_0	The observed differences are random
Alternative — H_1	The observed differences are reliable and caused by objective conditions

A statistical hypothesis testing is feasible only when the elements of the compared samples can generate a certain *value* (criterion), whose distribution law is known in case of correct H_0. Then, for this value, one can obtain a *confidence interval* into which with the given probability P_c its value falls. This interval is called a *critical domain*. If a value of the criterion falls into the critical domain, hypothesis H_0 is accepted. Otherwise, hypothesis H_1 is accepted.

In medical studies, $P_c = 0.95$ or $P_c = 0.99$ are used. *Significance levels* $\alpha = 0.05$ or $\alpha = 0.01$ correspond to these values.

For statistical hypotheses testing, a *significance level* (α) is the probability of rejecting the null hypothesis when it is true.

Please note that, as a matter of fact, hypotheses testing procedure is aimed at *detection of differences*, but not at finding a proof of their absence. When a value of the criterion oversteps the limits of the critical domain, we can say sincerely to the skeptic: "Well, what else do you expect?! If there had been no differences, the value rated would have been within the limits specified with the probability of 95% (or 99%). But it is not so at all!"

Well, if the value of the criterion falls into the critical domain, then there is no reason to believe that hypothesis H_0 is true. There is most likely one of the two possible causes.

1. Sample sizes are not large enough to detect differences. Continuation of the experiments is likely to be more successful.

2. There are differences. However, they are so small that they do not have practical importance. In this case, continuation of the experiments does not make sense.

Let us consider some statistical hypotheses applied to medical research.

§ 3.6. HYPOTHESES TESTING OF EQUALITY OF VARIANCES, FISHER F-CRITERION

In some clinical studies, the positive effect is proved not so much by a *value* of the parameter being studied as by its *stabilization*, reduction of its fluctuations. In this case, we encounter the problem of comparing two general variances according to the results of the sample examination. This problem can be solved by using *Fisher's criterion*.

Problem definition

There are two samples $\{X_1\}$ and $\{X_2\}$ extracted from general populations with *normal distribution law*. The sizes of samples are n_1 and n_2, and *sample variances* are s_1^2 *and* s_2^2 respectively. It is required to compare the *general variances*.

The hypotheses to test:

H_0 — general variances are *equal*;

H_1 — general variances are *different*.

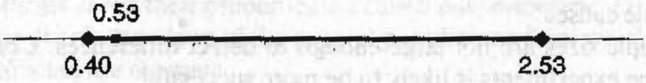
It is shown that if samples are extracted from the general populations with *normal distribution law*, then when hypothesis H_0 is true, the ratio of the sample variances is subject to the Fisher's distribution. So, the criterion for checking the validity of H_0 value F is used. It is calculated by the formula:

$$F = \frac{s_1^2}{s_2^2}, \tag{3.9}$$

where s_1^2 and s_2^2 are sample variances.

This relation is subject to Fisher's distribution with the numerator number of degrees of freedom $\nu_1 = n_1 - 1$, and the denominator number of degrees of freedom $\nu_2 = n_2 - 1$. The boundaries of the critical domain are found with tables of Fisher's distribution or using the computer function FINV.

For the example represented in Table 3.4, we obtain: $\nu_1 = \nu_2 = 20 - 1 = 19$; $F = 2.16/4.05 = 0.53$. For $\alpha = 0.05$ the boundaries of the critical domain are respectively: $F_{\text{left}} = 0.40$, $F_{\text{right}} = 2.53$.



The value of the criterion has fallen into the critical domain; so, hypothesis H_0 is accepted: the general variances of the samples are *equal*.

◆

§ 3.7. HYPOTHESES TESTING REGARDING EQUALITY OF AVERAGES, STUDENT'S *t*-CRITERION

The problem of comparing *averages* of two general populations arises when exactly the *value* of the studied indicator is of practical importance. For instance, when treatment time intervals by two different methods are compared, or the number of complications due to their administration are compared. In these cases, the Student's t-criterion can be used.

Problem definition

There are two samples $\{X_1\}$ and $\{X_2\}$ extracted from general populations with *normal distribution law* and *equal variances*. The sizes of the samples are n_1 and n_2, *the sample averages* are X_1 and X_2, and *the sample variances* are s_1 *and* s_2 respectively. It is required to compare *the general variances*.

The hypotheses being tested:

H_0 — general averages are equal;

H_1 — general averages are different.

It is shown that when hypothesis H_0 is true, value t, calculated by the formula:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(\nu_1 \cdot s_1^2 + \nu_2 \cdot s_2^2) \cdot (n_1 + n_2)}{(\nu_1 + \nu_2) \cdot n_1 \cdot n_2}}}, \tag{3.10}$$

is distributed according to Student's law with the number of degrees of freedom $\nu = \nu_1 + \nu_2 - 2$.

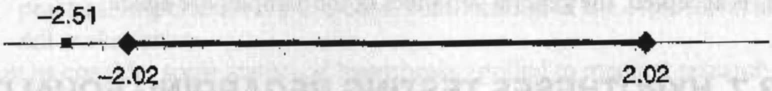
Here, $\nu_1 = n_1 - 1$ is the number of degrees of freedom for the first sample; $\nu_2 = n_2 - 1$ is the number of degrees of freedom for the other sample.

The boundaries of the critical domain are found with tables of t-distribution or using the computer function TDIST. The Student's distribution plot curve is symmetric with respect to zero, so the left and right boundaries of the critical domain are equal in their absolute value and opposite in their sign: $-t_b$ and t_b.

For the example represented in Table 3.4, we obtain:

$$\nu_1 = \nu_2 = 20 - 1 = 19; \quad \nu = 38,\ t = -2{,}51.$$

As $\alpha = 0.05$ $t_b = 2.02$.



The value of the criterion oversteps the left boundary of the critical domain, so we accept hypothesis H_1: general averages are *different*. At all that, the general population average for the *first sample* is SMALLER.

Applicability of Student's criterion

Student's criterion is applicable only to samples from *normal law* populations with *the same general variances*. If at least one of the conditions is violated, applicability of the criterion is doubtful. The requirement to a general population to be normal is usually ignored by referring to the *Central Limit Theorem*. Surely, the difference of the sample averages in the numerator of (3.10) can be considered normally distributed when $\nu > 30$. But the problem of equal variances is not subject to testing, and references to the fact that the Fisher criterion has not detected any differences cannot be taken into account. Nevertheless, the *t*-criterion is widely used for detection of differences in average values of general populations although without sufficient grounds.

The **nonparametric criterion**, which is used successfully for the same purposes and which does not require either *normality* or *equality of variances*, is considered below.

§ 3.8. NONPARAMETRIC COMPARISON OF TWO SAMPLES: MANN–WHITNEY TEST

Nonparametric tests are intended for detection of differences in distribution laws of two general populations. Tests that are sensitive to differences in general *averages* are called *slippage* test. Tests that are sensitive to differences in general variances are called *scale tests*. Mann–Whitney test is *slippage test*, and is used for detection of differences in average values of two general populations samples from which are represented in a *rank scale*. The indicators measured are placed on this scale in the ascending order, and then numbered with integers 1, 2... Exactly these numbers are called *ranks*. Equal values have the same rank. The value of the indicator is not important, but only the *ordinal number* that it takes among the other values.

In Table 3.5, the first group from Table 3.4 is expanded (line 1), ranked (line 2), and then the ranks of the same values are replaced by arithmetic mean values. For instance elements 4 and 4 being in the first line had been ranked 2 and 3, which were then replaced with the same values 2.5.

Table 3.5

3	4	4	5	5	5	6	6	6	6	6	6	6	7	7	7	7	8	8	9
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	2.5	2.5	5	5	5	10	10	10	10	10	10	10	15.5	15.5	15.5	15.5	18.5	18.5	20

Problem definition

Independent samples $\{X_1\}$ and $\{X_2\}$ are extracted from general populations with unknown distribution laws. The sizes of the samples are n_1 and n_2, respectively. Values of the sample elements are represented in the *rank scale*. One should check whether these general populations differ each from other.

Hypotheses being tested:

H_0 — the samples belong to the same general population;

H_1 —the samples belong to different general populations.

For checking such hypotheses, Mann–Whitney U-test is used.

First, a pooled sample $\{X\}$ is made up of two samples, elements of the former are ranked. Then the sum of ranks corresponding to the elements of the first sample is found. This sum is exactly the criterion for the hypotheses testing.

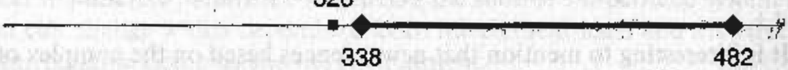
$$U = \text{the sum of the first sample ranks.} \tag{3.11}$$

For independent samples with sizes greater than 20, the value of U is subject to normal distribution, mathematical expectation and QD of which are equal to:

$$\mu = \frac{n_1 \cdot (n_1 + n_2 + 1)}{2}, \qquad \sigma = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}. \tag{3.12}$$

That is why, the boundaries of the critical domain are found with tables of normal distribution.

For the example represented in Table 3.4, we obtain: $\nu_1 = \nu_2 = 20 - 1 = 19$, $U = 339$, $\mu = 410$, $\sigma = 37$. For $\alpha = 0.05$ we obtain: $U_{\text{left}} = 338$, $U_{\text{right}} = 482$.



The value of the criterion oversteps the left boundary of the critical domain, so we accept hypothesis H1: the general populations have different distribution laws. For all that, the general population average for the *first sample* is SMALLER.

Chapter 4

Fundamentals of cybernetics

Cybernetics is the science of communications and automatic control systems.

The year when modern cybernetics was born is 1948, when the American mathematician N. Wiener published the work "Cybernetics: Or Control and Communication in the Animals and the Machine". Cybernetics studies common properties of various control systems irrespective of their material basis. These properties are manifested in nature, technology and in groups of people.

§ 4.1. CYBERNETICS AND OTHER SCIENCES

In general outline, the reader knows the subject of many natural, social and technical sciences, such as physics, mathematics, chemistry, biology, biophysics, history, electrical engineering, etc. Among these sciences, a special position is occupied by mathematics — a science studying spatial forms and quantitative relations of the real world. The exceptionality of this science is that it is an instrument of cognition in any field of human knowledge. All sciences, as it was already mentioned, develop, using mathematical laws to a certain extent. This is also true of cybernetics.

Wiener saw common problems and features in many different sciences. Control is carried out in the society, in many technical systems, in a living body. People, computers, and biological systems process information; it is transmitted through wire lines, radio channel, and neural structures.

Cybernetics has sprung up on the base of many sciences. It is not possible to enumerate everything, but the influence of technology, mathematics (theory of automated control, mathematical logic, theory of information and communication, computers, etc.) and physiology (theory of conditioned reflexes, principle of feedback, theory of functional systems, etc.) is undisputed.

The place of cybernetics in the system of sciences is sketchily shown in fig. 4.1.

It is interesting to mention that new sciences based on the complex of already existing ones continue to emerge nowadays. As an example, we can cite *synergetics*, which is the sphere of scientific research, the purpose of which is to reveal common laws in the processes of formation, stability, and destruction of ordered temporal and spatial structures in complex systems of different nature (physical, chemical, biological ones, etc.).

Many Russian and Soviet scientists have made direct or indirect contributions to development and creation of cybernetics. Among them, there are physiologists and physicians I.M. Sechenov (1829–1905), I.P. Pavlov (1849–1936), A.A. Bogdanov (1873–1928), P.K. Anokhin (1898-1974), V.V. Parin (1903–1971), N.M. Amosov (1913–2002), various technical specialists and mathematicians: I.A. Vyshnegradsky (1831–1895), A.M. Lyapunov (1857–1918), A.I. Berg (1893–1979), S.A. Lebedev (1902–1974), A.N. Kolmogorov (1903–1987), A.A. Kharkevich (1904–1965), V.A. Kotelnikov (1908–2005), L.V. Kantorovich (1912–1986), V.M. Glushkov (1923–1982), et al.

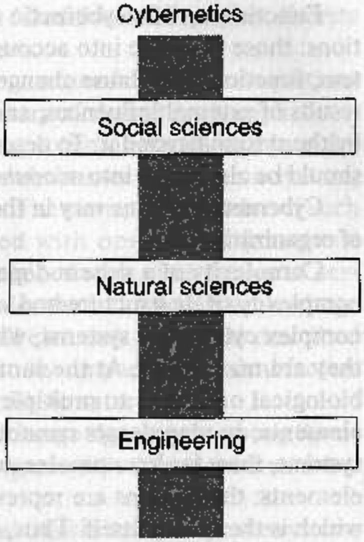


Fig. 4.1

§ 4.2. CYBERNETIC SYSTEMS

A cybernetic system is an ordered set of objects (elements of the system), interacting and interconnected, which are able to receive, memorize and process information, as well as exchange information.

Examples of cybernetic systems are groups of people, the brain, computers, and automatic machines. Respectively, objects of different physical nature like humans, cerebral cells, computer units can be elements of a cybernetic system etc.

The state of the elements of a system is described by a certain set of parameters, which are subdivided into continuous parameters assuming any real values in a certain interval, and discrete ones taking finite sets of values. For instance, the temperature of a human body is a continuous parameter, and its gender is a discrete parameter. In general, the state of a cybernetic system element can change which depends on both the element itself and the effect of the surrounding elements and the environment.

The structure of a cybernetic system is determined by the organization of relations between the elements of the system and is a function of the states of the elements and external influences.

Functioning of a cybernetic system is described by three families of functions: those that take into account changes in the state of elements of the system; functions that cause changes in the structure of the system, including the results of external influences, and functions that determine signals transmitted by the system beyond it. To describe a system more completely, its initial state should be also taken into account.

Cybernetic systems vary in their complexity, degree of definiteness and level of organization.

Complexity of a system depends on the number of elements that form it, complexity of its structure and variety of its internal connections. There exist complex cybernetic systems, which, however, can be known in detail, since they are man-made. At the same time, complex cybernetic systems; such as biological ones, due to multiple and undefined diverse interrelations of their elements, in many cases cannot be described in detail. In studying complex systems, there is also a process, which is opposite to division of the system into elements: the systems are represented in the form of enlarged units, each of which is the system itself. Thus, complex systems can consist of simpler ones. A higher-level system is a union of lower level subsystems, that is, the organization of a system is hierarchical.

Relationships can occur among levels of a hierarchy. The concept of elements itself in this sense is relative. In different cases, the same part of the system can be an element, a block, and the whole system. For instance, in studying cerebral functions, it can be considered as an element, while in studying the brain in connection with its internal structure, separate neurons should be taken as elements. In its turn, a neuron will be a cybernetic system when studying its cellular structure.

Cybernetic systems are divided into continuous and discrete ones. In continuous systems, all signals circulating in the system and the state of the elements are set by continuous parameters, in discrete system all such signals are set by discrete ones. However, there are also mixed (hybrid) systems, in which there are parameters of both types. Division of systems into continuous and discrete ones is conventional and is determined by the necessary degree of accuracy of the process studied, technical, and mathematical conveniences. Some processes or quantities of discrete nature, such as electric current (an electric charge discreteness: a charge cannot be less than the charge of an electron), it is convenient to describe with continuous quantities. In other cases, on the contrary, it makes sense to describe a continuous process with discrete parameters. For example, continuous renal excretory function can be conveniently described with a discrete five-point characteristic. In addition, in any physical measurements, when doing them at certain intervals of time, actually a set of discrete quantities is obtained. All of this confirms that discrete systems are more universal than continuous ones.

For studying of continuous systems, the apparatus of differential equations is used, for studying of discrete systems, the theory of algorithms is applied.

In cybernetics and technology, systems are conventionally divided into deterministic and probabilistic ones. ***A deterministic*** system is such a system, whose elements interact in a determined way. The state and behavior of such asystem is foreseen definitely and described with one-valued functions. Behavior of probabilistic systems can be determined with some degree of certainty since the elements of the system are under the influence of such a large number of influences that the interaction of all elements cannot be described accurately. An example is reaction of a body to the impact of physical factors like force, electricity, heat, etc.; it has a probabilistic nature.

The system is called closed if its elements exchange signals only among themselves. Unclosed, or open, systems exchange signals with the external environment.

To receive signals from the environment and transmit them into the system, every open system has receptors (***sensors or transducers***). In animals, as cybernetic systems, their receptors are the senses: touch, vision, hearing, etc., automatic machines have sensors: deformation gauges, photoelectric, inductive ones, etc. (see § 21.3).

Signals are transmitted to the environment by means of actuators called ***effectors***. Speech, hands, and facial expressions are effectors for humans as cybernetical systems. In a carbonated water vending machine the receptor is the button or coins receiver, the effector is the unit of carbonated water glassing.

Complex cybernetic systems have a characteristic property — an ability to accumulate information that can be used in subsequent operations of the control unit. This property is called memory, by analogy to a similar property of the human brain. Memorization in cybernetic systems is carried out in two ways: first, due to changes in the state of the elements of the system, and, second, because of changes in its structure.

§ 4.3. ELEMENTS OF INFORMATION THEORY

The central point of cybernetics is *information*. This term was mentioned in the course without special explanation as a commonly understood one. The word "information"[1] means, according to present-day ideas, a set of data and their transmission.

---

[1] *Informatio* (*Latin*) means clarification, notification.

Any phenomenon or event can serve as a source of information, but it should make sense and be a signal to this or that operation. Sometimes it is said that **information is a complex of data about the world around us that a human receives due to observation and communication with other people. People receive information when they feel pain, hunger, cold; see, hear, talk with other people, read books, etc.**

However, the idea that only human beings receive information is subjective. In fact, this concept has a broader meaning. Thus, **continuous regulation of viscera of animals and system of plant development is associated with transmission of information.**

We should not run to the other extreme, suggesting that any reflection of events in the world is information. It is hardly possible to consider that a drop in temperature in the mountains is information for the rocks about the onset of winter.

Transmission, reception, and processing of information are characteristic for systems that are complex enough in their structure; their peculiarity is the presence of control processes. The distinctive feature of information is its destroying any ignorance, reducing the uncertainty of a situation.

Scientific approach to the study of information was caused by the information explosion, an avalanche-like flow of information due to rapid development of science and technology in the middle of the 20th century.

The concept of information in cybernetics is of the same importance as the concept of energy and mass in physics. The section of cybernetics dealing with collection, transmission, storage, processing, and calculation of information was called the ***theory of information***. Let us consider briefly the elements of this theory.

Information is transmitted through communication channels in the form of ***signals*** generated by elements of a cybernetic system. ***A communication channel*** is a medium through which signals are transmitted. In a conversation, the signal is speech, and the communication channel is air; in music radio transmission, the signal is the sound, and communication channels are electromagnetic field and air.

Physical carriers of a signal can be various types of matter, which can alternate when transmitting the same signal. For instance, in radio transmission, a thought expressed orally is transmitted with bioelectric impulses to vocal muscles, causing their contraction, creating a sound image, which, due to membrane vibrations in a microphone, is converted into an electric pulse which is a signal transmitted over a distance. In this process, signals should meet the requirements of isomorphism. ***Isomorphism*** is such a correspondence of physically different phenomena, when the content of a message transmitted is preserved and not distorted.

Violation of isomorphism generates distortion of information. Distortion of signals because of isomorphism violation anddue to external interferences is called a ***noise***.

Depending on significance of the signals transmitted, they are divided into ***informative*** ones, carrying some information, and ***executory*** ones, that include any command to operate. ***Discrete*** and ***continuous*** signals are distinguished. An example of a discrete signal is the Morse code transmission or transmission of digits by current pulses; an example of a continuous signal is voltage change in a circuit corresponding to a temperature change.

Every message consists of a combination of simple signals of a certain physical nature. A complete set of such signals is called an *alphabet*, and a single signal is called a ***letter of the alphabet***. To transmit a message, it should be described using any alphabet, in other words, be encoded. ***Coding*** is a description of a message by a certain alphabet, i.e., establishment of a single-valued correspondence between the parameters characterizing the signal and information. Translation of this message into another alphabet is called ***conversion***, deciphering of the message is called ***decoding***.

To transmit messages in the economic and scientific life, encoding is performed by people. However, nature has created natural ways of encoding. These ways are of huge interest for science, like the way of encoding hereditary information about an adult multicellular body in a germ cell. Application of coding makes it possible to use a small alphabet to transmit a large amount of information. It turned out that any information can be encoded using two characters (0.1). This code is called ***binary***.

Transmission of any signal is associated with energy consumption, but the amount of information transmitted, and, even more so, its meaning does not depend on the signal energy. Moreover, very often a low-energy signal transmits a message that can trigger a process implying huge consumption of energy. For example, a nuclear explosion can be caused by pressing the switch button of the corresponding device; quiet information about someone's unsightly deed can cause an explosion of indignation.

In cybernetics, it does not matter what energy was consumed to transmit information, but it is essential how much information will be transmitted or can be transmitted through this or that communication channel. To quantify of information, you should abstract yourselves from the meaning of the message, similar to the way you abstract yourself from specific objects when doing a sum. For instance, adding two and three, we will obtain five regardless to what objects we are adding: apples, missiles, or stars.

Well, how is information quantified? It has already been noted that information makes sense when it reduces the degree of ignorance, that is the process of extracting information is associated with increasing certainty of our information about an object. The message carries information if from the totality of really possible events, some certain event is specified.

For example, reading a case history, a physician receives information about diseases of the patient: from a variety of different diseases, only those that this patient has had are mentioned in this document. A message about something already known carries no information: for a literate person, there is no information in the statement that after the 15th day of the month the 16th comes.

The more outcomes an event has, the more information about it a message carries. Thus, a single throwing of a dice (six faces) yields more information than throw a coin (two sides), because the first case has a greater number of equally possible outcomes than the other. They say that the amount of information varies inversely to probability.

Since probability is the measure of uncertainty of any events, it should be assumed that quantification of information is linked with basic concepts of the probability theory. Indeed, the modern method of information quantification is based on probabilistic approach when considering communication systems and message coding.

Let us consider a method of quantifying information contained in a single message offered by Shannon and used in modern information theory.

A measure of the amount of information can be found as a change in the degree of uncertainty of expectation of some event. Let us assume that there are k equally probable outcomes of an event. Then it is obvious that the degree of uncertainty of one event depends on k: in case of $k = 1$, the prediction of the event is certain, that is, the degree of uncertainty is zero; in case of large k, it is difficult to foresee an event; the degree of uncertainty is high.

Therefore, the desired function $f(k)$ (measure of quantity of information, or change in the degree of uncertainty) should be zero as $k = 1$ and should increase when k increases.

In addition, the function f should meet another condition. Let us assume that two independent experiments are conducted, one of them has k equally probable outcomes, and the other has l probable outcomes. It is natural to assume that the uncertainty of $f(kl)$ of joint appearance of some combination of events of the first and second experiment is more than $f(k)$ and $f(l)$ and is equal to the sum of the uncertainty of outcomes of each experiment:

$$f(kl) = f(k) + f(l). \tag{4.1}$$

The left part of the formula presents the function $f(kl)$ of the product kl, equal to the number of possible combination pairs of the outcomes of the first and second experiment. Logarithmic function $f(k) = \log_a k$ corresponds to formula (4.1).

$$\log_a kl = \log_a k + \log_a l. \tag{4.2}$$

In addition, the function obtained meets the following conditions: $\log_a 1 = 0$ and it increases with increasing k.

Since transition from one base of logarithms to another depending on the base is reduced to multiplication of the function $\log_a k$ by a constant factor, the base of logarithms is not of decisive importance and will only affect the choice of units of information amount.

So, function $\log_a k$ is considered to be a measure of uncertainty (amount of information) in case of k equally probable outcomes. The probability of each outcome is equal to $p = p_1 = p_2 = p_3 = \ldots = p_k = 1/k$. Since uncertainties of different events are added, the uncertainty of each individual outcome is equal to

$$\frac{1}{k}\log_a k = -\frac{1}{k}\log_a \frac{1}{k} = -p\log_a p. \tag{4.3}$$

In experiment with different probability outcomes $p_1, p_2, \ldots p_k$ the uncertainty measure for each individual outcome is expressed as follows (4.3):

$$-p_1 \log_a p_1, -p_2 \log_a p_2, \ldots - p_k \log_a p_k,$$

and the uncertainty measure for the whole experiment will be the sum of all the uncertainties:

$$H = -\sum_{i=1}^{k} p_i \log_a p_i. \tag{4.4}$$

This is the average value of the probability logarithm. By analogy with the Boltzmann formula [see (12.20)], H is called ***entropy*** or ***information entropy***. This value can be considered as a measure of information.

Searching the extremum of (4.4) we find that the event with equally probable outcomes has the highest uncertainty. The experiment in this case gives the most amount of information:

$$H_{\max} = -\left(\frac{1}{k}\log_a \frac{1}{k} + \frac{1}{k}\log_a \frac{1}{k}\right) + \ldots = \log_a k. \tag{4.5}$$

In the particular case of two equally possible events, the amount of information received from a message is equal to

$$H = -\left(\frac{1}{2}\log_a \frac{1}{2} + \frac{1}{2}\log_a \frac{1}{2}\right) = \log_a 2. \tag{4.6}$$

To choose a unit for the amount of information we assume that $a = 2$, then from (4.6) we have

$$H = \log_a 2 = 1.$$

This amount of information is taken as a bit (a bit is information contained in a message about one of two equally probable events). Assuming in (4.5) $a = 2$, we obtain that amount of information

$$H = \log_2 k \tag{4.7}$$

is expressed in bits.

Let us calculate the information obtained in case 1 falling out when throwing a dice. Using (4.7), we have

$$H = \log_2 6 \approx 2{,}6 \text{ bit}.$$

The concept of information is one of the most important in cybernetics as any control process is associated with receiving, accumulating, and transmitting information. Reflecting general properties of the material world, the concept of information serves as a philosophical category.

Information processes take place in the operation of any control systems — from transmission of hereditary characters to communication between people and machines. By analogy with the fact that in physics energy determines the measure of transformation of one impermanence form to another, in cybernetics information is a measure of processes of the material world reflection.

As already mentioned, information is transmitted through communication channels by signals. Information received from the source by receiving elements (sense organs, microphones, photocells, etc.) is converted by the encoder into a form convenient for transmission of the signal: an electric signal, and is transmitted through a communication channel to a receiver, where the information is decoded, e.g., into sound, and communicated to the listener. General block diagram of an information transmission system is shown in fig. 4.2.

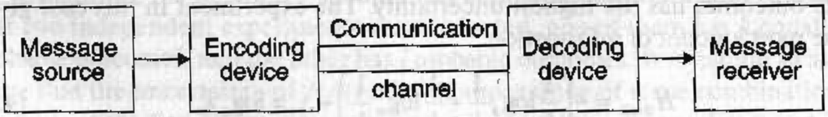


Fig. 4.2

In conclusion, please note that some quantitative expressions of the information theory have not yet found their application to medical cybernetics. This circumstance is due to the general, still to a great extent qualitative, nature of medicine.

§ 4.4. CONTROL AND REGULATION

In order to ensure a purposeful changing of the cybernetic system behavior, a control is needed.

A control is influencing a cybernetic system (object) in accordance with the existing program or the purpose of its operation. In short, control is action on an object to achieve a set goal.

Control goals can be different. In the simplest case, it is simply maintaining some parameter constant (constant humidity or temperature in a room). In more complicated cybernetic systems, the purpose of control is adaptation to changing conditions, such as adapting to the changing habitat of a biological individual.

It was established that the scheme of controling objects of different nature is common both for the organic world, including the mechanisms of control in a living body and mechanisms of biological evolution, and for the inorganic world, up to computers and spaceships control.

This similarity makes it possible to draw analogies between living systems which got refined in the long process of evolution, and technical devices that are simpler and less perfect.

Studying biological control systems and their comparison with technical systems, on the one hand, makes it possible to find new principles for creation of more complex technical devices, and on the other hand, to understand control principles underlying biological objects and processes. The first side of the problem is the content of a research direction called bionics.

In any control system it is necessary to distinguish the control unit and the control object, and also communication lines (communication channels) between them. The control unit is a quite important part of a cybernetic system. A control system processes information received and generates controlling impacts. Information processing exists in various natural and artificial control systems. These include thinking, information processing in computer-aided systems, changes in hereditary information in the process of evolution of biological species, etc. Controlling operations are transmitted by appropriate effectors onto a control object. Physical processes that transfer information and are a signal carry out communication. Having received a signal, a control object will pass to the corresponding state.

Of the greatest interest is a control in which operations ensureing achievement of the control goal given are performed by a system that operates without human intervention in accordance with a predetermined algorithm. This variant is called *automatic control*.

A variety of automatic control is ***automatic regulation***. This term is used in cases where the purpose of control is automatic maintenance of constancy

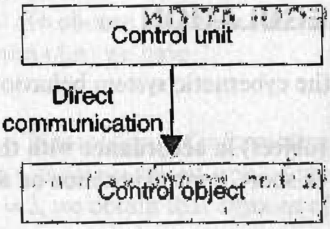


Fig. 4.3

or change of a physical parameter of the object of control (regulation) according to the law required. In this case, the controlling unit can be called *regulator*.

If a control system does not receive or take into account information from the control object, the system is called an *open-loop system*. Such control is sketchily shown in fig. 4.3 indicating a channel (line) of direct communication. Such control is implemented in traffic lights, genetic system, and computers.

In open-loop system mode, the type of control implemented is referred to as disturbance-stimulated control. Let us explain this on an example of a device that automatically maintains comfortable temperature in a room (fig. 4.4). In this, the controlled object is the air conditioner. A disturbance (outdoor temperature) affects the regulator (special thermometer) and influences the air temperature in the room. Depending on disturbances the thermometer sends signals to the air conditioner to switch to heating or cooling mode. The air of the appropriate temperature enters the room. It is essential that in this system heating or cooling of the air in the room depends on the ambient temperature rather than on the room temperature.

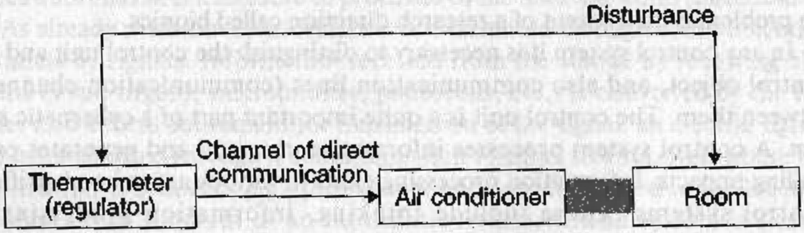


Fig. 4.4

More common and effective are control systems with feedback that are called closed-loop control systems (fig. 4.5). In this case, the control unit processes information received from outside both from other objects of the system and from the control object trough the feedback communication line.

Feedback is transmission of impact* or *information from the output of the system (element) to its input, in particular, the influence of the control object on the control unit.

Positive and negative feedback are differed. In case of positive feedback, the results of a process tend to strengthen it. In technical devices, positive feedback

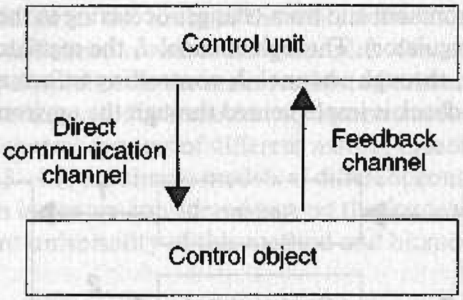


Fig. 4.5

contributes to the system's transition to another equilibrium state or causes an avalanche process.

Negative feedback prevents any development, change of a process and stabilizes it. Negative feedback is used in closed-loop control systems.

As a technical system with a negative feedback, let us consider a thermostat temperature controller, which uses a contact thermometer (fig. 4.6).

As the temperature drops below the set point, the mercury column in the thermometer breaks the contact in the relay circuit, the relay turns on the heater, and the temperature rises. At temperatures above normal, the mercury column closes the relay circuit and the heater is switched off. The considered system makes it possible to maintain the thermostat temperature in a certain interval. This example illustrates automatic disturbance-stimulated control (regulation).

Self-governing (self-regulating) ***systems*** are among cybernetic systems with negative feedback (closed-loop control system). A self-governing system is, for instance, the body of an animal, in which a constant composition of blood, temperature and other parameters are autonomously maintained. A system consisting of a group of animals and predators eating them, such as hares and wolves, is also self-regulating. An increase in wolf population generates food deficiency (hares) which in turn generates reduction in the number of wolves, hence it increases the hares' number, etc. As a result, if we ignore other factors (wolf culling, drought, etc.), the number of wolves and hares is maintained in this system at a certain level.

The diagram of self-governing system of this type can be represented as consisting of the following parts (fig. 4.7): control object, which affects the environment, a sensor that sensor receives information

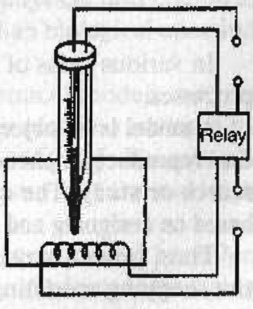


Fig. 4.6

both from the environment and from changes occurring in the controlled object, and control unit (regulator). Through channel *1*, the regulator receives primary input information, through channel *2*, controlling information is sent to the control object. Feedback is implemented through the environment and the sensor.

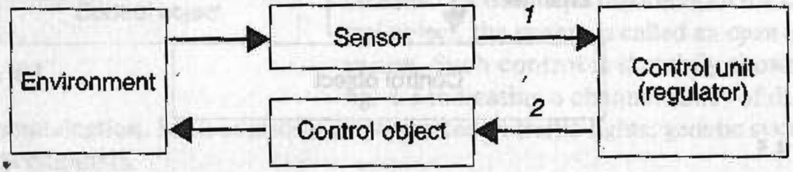


Fig. 4.7

Studying self-governing systems is of particular interest for physiology and biology.

There are systems of optimal control, whose purpose is to maintain the extremal (minimum or maximum) value of a certain parameter depending on external conditions and control signals from the system.

The simplest example of such regulation is operation of an air conditioner that maintains the temperature in accordance with air humidity. An optimal control system is also appropriate in cases when the function of a system is reduced to maintaining parameters regulated to have maximum or minimum values upon a change in nonregulated parameters.

Issues of control are discussed in more detail in the special theory of control systems. The main underlying principles are feedback and multistage control. Feedback makes it possible for a cybernetic system to take into account the real circumstances and coordinate them with the desired behavior. A multistage control structure ensures reliability and stability of cybernetic systems.

§ 4.5. SIMULATION

In various areas of knowledge, models are used to study real systems and processes.

A model is an object of any nature, speculative or materially implemented, that reproduces a phenomenon, process, or system with the purpose of their research or study. The method of studying phenomena, processes and systems, based on designing and studying their models, is called simulation.

Thus, present-day understanding of simulation includes not only a subjective, copying modeling, like creation of a glider model, but also a scientific method of research and cognition of essence of phenomena and objects.

The basis of simulation is the unity of the material world and the attributes of matter — space and time, as well as the principles of movement of matter.

In cybernetics, simulation is the main method of scientific cognition. This is due to the abstract nature of cybernetics, common structure shared by cybernetic systems and control systems of different nature. Essentially, the diagrams presented in fig. 4.3–4.7, are simple models of different control systems. In this section, simulation issues are considered beyond the framework of cybernetics, taking into account universality of this method and biomedical nature of the reader's interest.

Let us focus on the main, most essential types of models: ***geometric***, ***biological***, ***physical*** (***physicochemical***), and ***mathematical*** ones.

Geometric models are the simplest type. This is copying the external shape of theoriginal. Plaster casts used in teaching anatomy, biology and physiology are geometric models. In everyday life, geometric models are often used for cognitive or decorative and entertainment purposes (models of cars, railways, buildings, dolls, etc.).

Development of biological (physiological) models is based on reproducting certain states, such as disease of experimental animals in laboratory conditions. In the experiment, mechanisms of the state onset, its course, methods of influencing the body for its change are studied. Such models include artificially induced infections, hypertrophy of organs, genetic disorders, malignant neoplasms, artificially induced neuroses and various emotional states.

To create these models, the experimental body is exposed to a variety of impacts: infecting with microbes, injection of hormones, changes in the composition of food, impact on the peripheral nervous system, changes in conditions and habitat, etc.

Biological models are important for biology, physiology, pharmacology, and genetics.

Creation of physical and physicochemical models is based on reproduction of biological structures, functions, or processes with physical and chemical methods. Physicochemical models are more idealized than biological ones and are a far likeness of biological object being simulated.

As an example of one of the first physical and chemical models, we can give the model of living cell growth (1867), in which growth was imitated by growing crystals of $CuSO_4$ in water solution of $Cu[Fe(CN)_6]$. This simple model is based only on external, mainly qualitative, likeness of the model to the original.

Models based on quantitative similarity are more complex and are often based on the principles of electrical engineering and electronics using experimental material from electrophysiology.

Models developed are used in mechanical engineering with electronic control, simulating some acts of animal behavior (formation of a conditioned reflex, memory, inhibition, etc.). For a demonstrative effect, these machines are often given the appearance of animals: mice, turtles, and squirrels.

It is of practical importance to simulate physical and chemical living conditions of individual cells, organs or the whole body. Solutions artificially developed imitate a nutrient medium for existence of separate organs and cells outside the body.

Artificial biological membranes make it possible to study the physical and chemical nature of their ion permeability and the effect of various external factors on it.

Mathematical simulation of biological objects is an analytical description of idealized processes and systems adequate to real ones.

There are no ideal systems and processes in nature, but, within certain limits, results obtained can be applied to real processes and systems since they have properties common with ideal ones. This method of abstraction is used in physics too.

Mathematical models are built either based on experimental data (material, or subject simulation), or speculatively, using a hypothesis or a regularity known of any phenomenon. At the same time, theoretical simulation requires subsequent experimental testing.

Theoretical simulation is especially useful where experimentation is impossible or difficult.

Running a mathematical model of a biological process, which is difficult to reproduce in the experiment, on the computer, makes it possible to foresee changes of the process depending on conditions and some new phenomena. For instance, study of the cardiac activity model based on the theory of relaxation oscillations, has made it possible to foresee a special cardiac rhythm disturbance, subsequently detected in humans.

In a number of cases, the same differential equations are obtained for physically different models. For example, the same equations describe damped mechanical [see (7.33)] and electrical [see (18.13)] oscillations or the aperiodic discharge of a capacitor [see (18.17)], absorption of light by a material [see (29.6)] and the law of radioactive decay [see (32.8)]. In this similarity of differential equations relating to different phenomena, the unity of nature can be seen. Such a peculiarity makes it possible to use analogies for mathematical simulation, and the corresponding models are called ***object-mathematical models of direct analogy.***

Studying phenomena by means of mathematical models is divided into four stages.

The first stage consists in selecting simulation objects and formulating laws connecting them. It ends with a mathematical representation of ideas of the relations among the objects of the model in a written form.

At the second stage, mathematical problems generated by the mathematical model are studied. The purpose of this stage is to solve the primal problem, that is to obtain data that can be compared with the results of experiments or observations. To solve these problems, mathematical apparatus and computer technology are used to obtain quantitative information.

The third stage makes it possible to find out how the proposed model meets practice as the criterion of truth. Solution of this problem is associated with how theoretical conclusions correspond to the experimental results. At this stage, the inverse problem is often solved, when some previously unknown characteristics of the model are determined by comparing output information with the results of observations.

The proposed model is wrong if no values of its parameters it is possible warrant coordinating the output information with the experiment.

The fourth stage includes analysis of the model after accumulation of data about it, and its further modernization.

Depending on the nature of models, they are conventionally divided into phenomenological and structural ones.

Phenomenological (functional) models reflect temporal and cause-and-effect relationships among the parameters characterizing functions of a biological object without taking its structure into account.

The object is considered a "black box", a system in which only input and output values are available to an external observer, and its internal structure is unknown (fig. 4.8). "Black box" method is widely used to solve problems of simulating complex cybernetic systems where the behavior of a system is of interest. For instance, taking in account the complex structure of the human brain and the risk associated with direct instrumental invasion into its structures, it is reasonable to study the brain as a "black box". This can be done researching the mental capacity of a person, his responses to sound, light, etc.

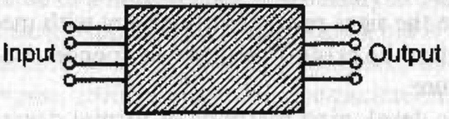


Fig. 4.8

Structural models are made using object's structure reflecting its hierarchical levels.

For all that, partial functions of individual subsystems are included into the structure. Such models reflect the essence of biological systems better, but they are difficult for quantification.

Creation of models is done according to a certain plan. First, the purpose of simulation is formulated, and then the hypothesis representing a qualitative description of the system, the type of model and mathematical methods of its description are chosen depending on the purpose and type of information.

The final stage is model creation and its comparison with the object system for identification.

§ 4.6. CONCEPT OF BIOLOGICAL AND MEDICAL CYBERNETICS

Biological cybernetics is an area of research in which ideas, methods and technical facilities of cybernetics are applied to problems of biology and physiology.

Biological cybernetics can be represented by their theoretical and practical parts. The main problem of theoretical biological cybernetics is studying general problems of control, storage, processing and transmission of information in living systems. One of the most important methods of practical biological cybernetics is the method of simulation, i.e., simulation of the structure and behavior of biological systems. Developing this method, biological cybernetics also includes design of artificial systems that reproduce the activity of individual organs, their internal connections and external interactions. In this relation, biological cybernetics links up with medical cybernetics.

Medical cybernetics is an area of research connected with the use of ideas, methods and technical facilities of cybernetics in medicine and public health. Medical cybernetics can be conventionally divided into the following groups.

1. ***Computational diagnostics of diseases.*** This group is mainly connected with use of computers for making diagnoses.

The structure of any diagnostic system consists of medical memory (cumulative medical experience in certain diseases) and a logical unit that make it possible to compare the signs revealed in a patient with medical interview and laboratory tests against the existing medical experience. A diagnostic computer has the same structure.

The first step is developing methods of formal description of patient's health; they conduct a thorough analysis to clarify clinical parameters and indicators used for diagnostics. Indicators that lend themselevs to quantification are mainly selected.

In addition to quantifying physiological, biochemical and other characteristics of a patient, computational diagnostics requires data about the occurrence rate (a priori probability) of clinical syndromes and diagnostic indicators, their classification, interdependence, evaluation of diagnostic effectiveness of indicators, etc. All these data are stored in the memory of the computer.

The next step is choosing an algorithm. The computer compares the patient's manifestations with data stored in its memory.

The logic of computational diagnostics corresponds to the logic of a physician making a diagnosis: the set of signs is compared with previous medical experience.

A computer does not diagnose a new (unknown) disease. A physician having found an unknown disease will be able to describe its indicators. The details of such a disease can be revealed only by special studies. Computers in such studies can be of an auxiliary importance.

2. ***Cybernetic approach to treatment.*** Having made the diagnosis, the physician administers treatment, which, as a rule, is not reduced to a single procedure. This is a complex process, during which the physician repeatedly receives medical, biological information about the patient, analyzes this information, and in accordance with it clarifies, changes, stops or continues the therapy.

Cybernetic systems are characterized by a purposeful influence of the control system on a controlled object (see § 4.4).

The physician controls the patient; the system "physician–patient" is cybernetic; so cybernetic approach is applicable to management of disease too. However, despite such opportunities, introduction of innovation, ideas, methods and technical facilities of cybernetics in this most important part of medicine, is still limited.

Nowadays, cybernetic approach to management of disease facilitates the work of a physician, makes it possible to administer more effective management of serious patients, as well as to take opportune measures for complications during surgery, develop and monitor pharmaceutical therapy, and create bio-controlled prosthetic appliances.

Let us consider the applicability of this approach briefly.

Controling the state of a human body is necessary in many fields of human activity (sports, industry, education, military sphere), but it is especially important in stressful situations or in such medical conditions as surgery with use of cardiopulmonary bypass, artificial respiration, resuscitation, anaesthesia, etc.

For the foregoing purposes, ***information systems of operational medical control*** (ISOMC) are created; they do reading biomedical information, automatic recognition of patients' functional status, recording disorders in the body functioning, diagnostics of diseases, controlling devices, which regulate essential body functions.

Problems of operational medical control include monitoring the actual state of critical patients using monitoring systems, monitoring the state of health in extreme conditions (stress, weightlessness, hyperbaric conditions, environment with low oxygen content, etc.).

Implementation of intensive therapy principle was made possible due to development of a complex that permits monitoring the actual patient's state automatically and continuously and inform about its changes.

It is especially important to obtain quick and accurate information about the patient's state during surgery. In this case, a huge number (about 1,000) of different parameters characterizing the patient's actual state is recorded. It is almost impossible for a physician to analyze and monitor so many parameters within an extremely short time interval. In such cases, computers help, especially in entering the available medical records, information about the type and dosage of medications, instructions of measures to be taken in critical situations.

General data of patients being operated are input into the computer preliminarily. Data about the actual state are input from the time of bringing the patient to the operating room. In addition to information about the patient's actual state, information about the time of administration, type, and dosage of anesthetics and medications is entered, and continuous recording of medical and biological parameters begins. As a result, if any parameters are beyond critical values, the computer will report in the form of sound or light signals about danger, will give information to the recording device, explaining the reasons for giving alarm, and recommendations for their elimination.

One more possibility of using cybernetics in medicine is mathematical simulation of management, which can serve as the basis for calculating optimal therapeutic effects. Thus, it is possible to compute the process of medication administration into patient's body in order to bring about the best possible effect.

Cybernetic approach is implemented in creation of complex prosthetic appliances that replace some organs. Let us explain this with an example.

Research of myoelectricity showed that since they can be read directly on the muscles, one can determine the information sent to the muscles (executive, controlled units) by the central nervous system (control system). It has also been found that myoelectricity can be generated in a muscle due to the central nervous system commands without their execution, e.g., if a limb or its part is absent.

These properties of myoelectricity have made it possible to develop active prosthetic appliances of limbs. A usual prosthetic appliance, like that of a leg, used to recover only a part of the function, which is support; control and coordination function were lost.

Prosthetic appliances for limbs with myoelectric control were developed. To control such limbs, special systems were developed; they include devices for reading action potentials, amplifier and converter that amplifies the signal and transforms it into a form suitable to control the mechanical part of the prosthetic appliance (electric motors, reduction gears, etc.) and driving the prosthetic appliance itself (hand, fingers, foot, etc.).

Using transducers (sensors) that receive external influences on an artificial organ, feedback is achieved: an electrical signal from the transducer is transformed into a signal, which is similar to current pulses in sensory nerves of a living body; it is sent from the periphery to the center through the intact areas of the skin of a bad limb.

3. ***Automated control systems and possibilities of their application in public health.*** In the previous sections, the focus was on control processes in biological systems. However, in its original version the term "control" was more associated with the concept of "management" and referred to management of the economy, enterprise, i.e., a team of people who perform a specific task. Such understanding of management is, surely, also cybernetic and therefore the management process can be optimized using methods and technical facilities of cybernetics.

Such optimization led to creation of automated control systems (ACS). ACS differs from traditional forms of management in their wide use of computers for acquisition and processing of information, as well as new organizational principles for implementation of the most effecient management of the corresponding object (system).

Control objects of ACS differ both in their scale and in their purpose: workshop sector, physician's office, admission room, enterprise, school, hospital, public health, industry, national economy, etc.

Depending on the level of hierarchy, ACS are divided into separate systems. For instance, in almost any branch of economy, it is possible to identify a ***branch automated control system*** (BACS).

***Public Health* is a sector of Russian economy; therefore, the BACS "Zdravookhraneniye" (Public Health, Rus.) was developed to manage this branch.**

Without going into details of such BACS, which is a problem of a special course in a medical university, we note only some of its distinctive features.

Any BACS can be built on the basis of models that take into account not only connections within a branch, but also inter-branch connections, i.e., the relationship of this or that system with the entire national economy. As applied to the BACS "Zdravookhraneniye", the model should include both the control unit and other elements: prevention, treatment (with diagnostics), medical research, personnel, and procurement.

Each of the foregoing elements (units) of the BACS is connected with elements of the same system, as well as with other systems. Let us illustrate this with an example of disease prevention. It includes immunization, screenings, health education, etc. Screenings depend on availability of trained medical staff, provision with equipment, etc. (internal connections and dependencies), the state and development of industrial enterprises, distribution of inhabitants in geographical areas, etc. (external connections beyond this BACS).

The priorities of the BACS "Zdravookhraneniye" include automation of acquisition and analysis of statistical information in the main areas of medical activities and resolution of optimization problems of some control processes.

Section 2

MECHANICS. ACOUSTICS

Mechanics is a part of physics in which mechanical motion of material bodies is studied. Mechanical motion means change in the position of a body or its parts in space over time.

Mechanics based on Newton's laws is called classical mechanics. It deals with motion of macroscopic bodies at velocities much less than the velocity of light in free space.

The subject matter of this section can be of interest for the following reasons:

- understanding mechanics of motion of the whole body for purposes of sports and space medicine, mechanics of a human musculoskeletal
- system for purposes of anatomy and physiology;
- learning about mechanical properties of biological tissues and fluids;
- understanding the ear and vestibular apparatus operation as physical devices, the heart as a pump, etc.;
- explanation of the biophysical mechanism of action of ultrasound;
- understanding physical fundamentals of some laboratory techniques used in practice of biomedical research, e.g., centrifugation.

Chapter 5

Mechanics of rotational motion

Observing complex movements, such as the movement of a human body (walking, running, jumping, etc.), it seems difficult or even impossible to describe motion of all its points. However, analyzing such movements, it can be seen that they consist of simpler ones — translational and rotational movements.

The reader already knows the mechanics of translational motion, so the section begins with consideration of rotational motion. The simplest one is rotation of a solid body about the fixed axis. This case makes it possible to familiarize yourself with the specific character, terminology, and laws of rotational motion.

§ 5.1. KINEMATICS OF ROTATIONAL MOTION OF A PERFECTLY RIGID BODY ABOUT A FIXED AXIS

A perfectly rigid body is such a body for which the distance between any two points does not change.

The size and shape of a perfectly rigid body do not change when it moves.

The concept of "perfectly rigid body" is a physical abstraction since any kind of a body is exposed to deformations. However, in many cases deformation can be neglected.

The simplest case of rotational motion of a perfectly rigid body is rotation relative to a fixed axis. This is a motion in which the points of the body move in circumferences, the centers of which lie on the straight line, called the axis of rotation.

It is known that in some cases it is not necessary to specify the movement of all its points; in translational motion, it is sufficient to specify the movement of any one point of the body.

For rotational motion about the axis, the points of the body move along different trajectories, but at the same time all the points and the body itself rotates through the same angle. To characterize the rotation, let us draw in a plane perpendicular to the axis, the radius vector r_i to some point i (fig. 5.1). **Time dependence of angle α of rotation of the radius vector relative to a certain selected direction OX is the equation of rotational motion of a solid body around a fixed axis:**

$$\alpha = f(t). \tag{5.1}$$

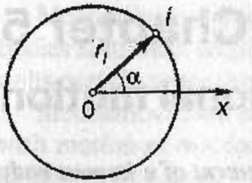


Fig. 5.1

The rotation rate of the body is characterized by angular velocity equal to the first derivative of the rotation angle of the radius vector by time:

$$\omega = \mathrm{d}\alpha/\mathrm{d}t. \tag{5.2}$$

Angular velocity is a vector that is directed along the axis of rotation and depends on the direction of rotation according to the right-hand screw rule (fig. 5.2). An angular velocity vector, in contradistinction to velocity and force vectors, is a sliding vector: it does not have a specific point of application, and it can be located anywhere on the axis of rotation. Thus, assigning the vector ω indicates the position of rotation axis, direction of rotation, and absolute value of angular velocity.

The rate of angular velocity change is characterized by angular acceleration equal to the first derivative of angular velocity by time:

$$\varepsilon = \mathrm{d}\omega/\mathrm{d}t \tag{5.3}$$

or in the vector form:

$$\vec{\varepsilon} = \mathrm{d}\vec{\omega}/\mathrm{d}t. \tag{5.4}$$

We can see from (5.4) that an angular acceleration vector coincides in direction with the elementary, rather small change in the angular velocity vector $\mathrm{d}\vec{\omega}$: in case of accelerated rotation angular acceleration is directed in the same way as angular velocity, in case of decelerated rotation it is opposite to the velocity.

Since angular displacement of all points of a perfectly perfectly rigid body is the same, then, according to (5.2) and (5.3), at the same time, all points of the body have the same angular velocity and the same angular acceleration. Linear characteristics, such as displacement, velocity, and acceleration, are different for different points. Let us show a relation, which can be derived independently, in the scalar form, between the linear and angular characteristics for *the i-th* point moving along the circumference with radius r_i:

$$\Delta s_i = r_i \Delta\alpha; \quad \upsilon_i = r_i \omega;$$

$$a_{ni} = \upsilon_i^2/r_i = r_i \omega^2; \quad a_{\tau i} = r_i \varepsilon; \quad a_i = r_i \sqrt{\omega^2 + \varepsilon^2};$$

where $\Delta\alpha$ is the rotation angle; Δs_i is the distance passed by the material point; υ_i is the point velocity (linear velocity); a_i is acceleration of the point; a_{ni} is normal and $a_{\tau i}$ is tangential component of acceleration (fig. 5.3).

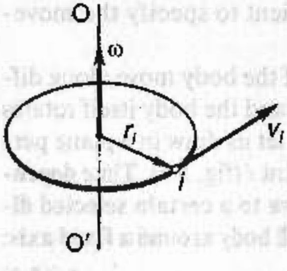


Fig. 5.2

In conclusion, we represent formulas for kinematics of rotational motion of a perfectly rigid body about the fixed axis obtained by integrating the corresponding expressions:

equation for uniform rotational motion [see (5.2)]:

$$\alpha = \omega t + \alpha_0 \tag{5.5}$$

(α_0 — initial value of the angle);

dependence of angular velocity on time for uniformly variable rotational motion [see (5.3)]:

$$\omega = \varepsilon t + \omega_0 \tag{5.6}$$

(ω_0 — initial angular velocity);

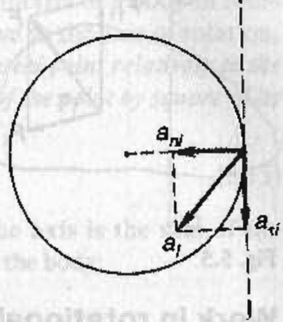


Fig. 5.3

equation for uniformly variable rotational motion [see (5.1) and (5.6)]:

$$\alpha = (\varepsilon t^2/2) + \omega_0 t + \alpha_0. \tag{5.7}$$

It is expedient to compare these formulas with similar dependencies for translational motion.

§ 5.2. BASIC CONCEPTS. EQUATION FOR GYRODYNAMICS

Moment of force

May force $\overline{F}_i$ be applied to some point i of a perfectly rigid body; it lies in a plane which is perpendicular to the axis of rotation (fig. 5.4).

Moment of force relative to the axis of rotation is the vector product of the radius vector of point *i* by the force:

$$\overline{M}_i = \overline{r}_i \times \overline{F}_i. \tag{5.8}$$

Expanding it, it is possible to write:

$$M_i = F_i r_i \sin\beta, \tag{5.9}$$

where β is the angle between vectors $\overline{r}_i$ and $\overline{F}_i$. Since the arm of force $h_i = r_i \sin\beta$ (see fig. 5.4), then

$$M_i = F_i h_i. \tag{5.10}$$

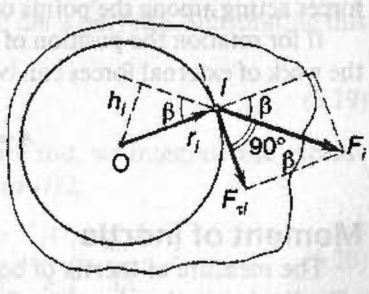


Fig. 5.4

If the force acts at an angle α to the plane of rotation (fig. 5.5), then it can be decomposed into two components. One

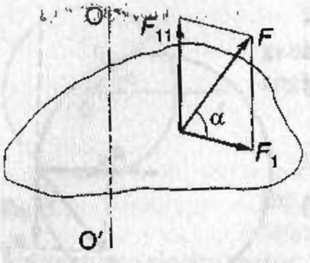


Fig. 5.5

of them lies in a plane which is perpendicular to the axis of rotation, and the other is parallel to this axis and does not affect the rotation of the body (in the real case it acts only on the bearings). Hereafter, only forces lying in a plane, which is perpendicular to the axis of rotation, will be considered.

Work in rotational motion

Let a body move by a sufficiently small angle dα under the action of force F_i (see fig. 5.4). Let us find the work of this force.

The expression known from secondary school for the work of force in this case should be written as:

$$dA_i = F_i \cos(90° - \beta)\, ds_i, \qquad (5.11)$$

where dA_i is elementary work of force F_i, i.e., the work corresponding to a sufficiently small angle of rotation; ds_i is elementary displacement of a material point for this rotation (sufficiently small part of an arc). Taking into account that $F_i \cos(90° - \beta) = F_i \sin\beta$ and $ds_i = r_i\, d\alpha$, and the relation (5.9), from (5.11) we have

$$dA_i = F_i \sin\beta\, r_i\, d\alpha = M_i\, d\alpha. \qquad (5.12)$$

So then, **elementary work of force for rotational motion is equal to the product of the moment of force by the elementary angle of rotation of the body.**

If several forces act on the body, elementary work done by all of them is determined similarly (5.12):

$$dA = M d\alpha, \qquad (5.13)$$

where M is the total moment of all external forces acting on the body.

We recommend you to prove by yourself that the total work of all internal forces acting among the points of a perfectly rigid body is equal to zero.

If for rotation the position of the radius vector changes from α_1 to α_2, then the work of external forces can be found by integrating expression (5.13):

$$A = \int_{\alpha_1}^{\alpha_2} M\, da. \qquad (5.14)$$

Moment of inertia

The measure of inertia of bodies in translational motion is mass. Inertia of bodies in rotational motion depends not only on mass, but also on its distribution in space relative to the axis. The measure of inertia of a body in rotation is characterized by the moment of inertia relative to the axis of rotation. Let us first note that the ***moment of inertia of a material point relatively to the axis of rotation is a value equal to the product of mass of the point by square of its distance from the axis***:

$$J_i = m_i r_i^2. \tag{5.15}$$

The moment of inertia of the body relative to the axis is the sum of the moments of inertia of all material points that make up the body:

$$J_i = \sum_{i=1}^{T} m_i r_i^2. \tag{5.16}$$

The moment of inertia of a continuous body is usually determined by integration:

$$J = \int\limits_{\substack{\text{In total}\\ \text{volume}}} r^2\,\mathrm{d}m. \tag{5.17}$$

As an example, let us derive the formula of ***the moment of inertia of a thin uniform rod*** l long and of mass m relative to an axis perpendicular to the rod and passing through its middle (fig. 5.6). Let us choose a sufficiently small part $\mathrm{d}x$ long with mass $\mathrm{d}m$ at a distance x from axis OO′. Due to the smallness of this part, it can be taken as a material point and its moment of inertia [see (5.15)] is equal to:

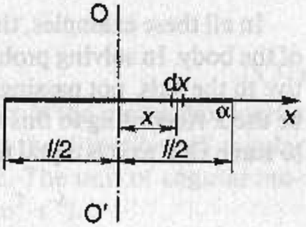


Fig. 5.6

$$\mathrm{d}J = x^2\,\mathrm{d}m. \tag{5.18}$$

The mass of the elementary part is equal to the product of linear density m/l by the length of the elementary section: $\mathrm{d}m = (m/l)\,\mathrm{d}x$. Substituting this expression into (5.18), we obtain

$$\mathrm{d}J = (m/l)\cdot 2\mathrm{d}x. \tag{5.19}$$

To find the moment of inertia of the entire rod, we integrate the expression (5.19) over the entire rod, i.e., from $-l/2$ to $+l/2$:

$$J = \frac{m}{l}\int_{-l/2}^{+l/2} x^2 dx = \frac{m}{l}\frac{x^3}{3}\bigg|_{-l/2}^{+l/2} = \frac{m}{3l}\left(\frac{l^3}{8} + \frac{l^3}{8}\right) = \frac{ml^3}{12}. \tag{5.20}$$

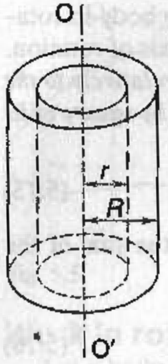


Fig. 5.7

Let us show expressions for the moments of inertia of different symmetric bodies with mass m:

- ***homogenous hollow cylinder*** (hoop) with internal radius r and external radius R relative to axis OO′, coinciding with the geometric axis of the cylinder (fig. 5.7):

$$J = m(r^2 + R^2)/2; \tag{5.21}$$

- ***thin wall cylinder*** ($R \approx r$) or ring [see (5.21)]:

$$J = mR^2; \tag{5.22}$$

- ***solid homogenous*** cylinder ($r = 0$) or disk [see (5.21)]:

$$J = mR^2/2; \tag{5.23}$$

- ***uniform sphere*** relative to the axis passing through its center:

$$J = 2/5mR^2; \tag{5.24}$$

- ***rectangular parallelepiped*** relative to axis OO′, passing through its center perpendicular to the plane of the base (fig. 5.8):

$$J = m(a^2 + b^2)/12. \tag{5.25}$$

In all these examples, the axis of rotation passes through the center of mass of the body. In solving problems determining the body moment of inertia relative to the axis, not passing through the center of mass, Huygens theorem can be used. According to this theorem, the moment of inertia of the body relative to some OO′ axis is equal to:

$$J = J_0 + md^2, \tag{5.26}$$

where J_0 is the moment of inertia relative to the parallel axis passing through the center of mass OO′; m is the body mass; d is the distance between two parallel axes (fig. 5.9). The unit of moment of inertia is a *kilogram-meter squared* (kg · m^2).

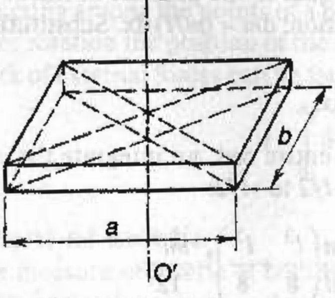


Fig. 5.8

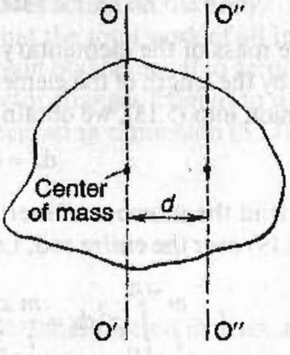


Fig. 5.9

Angular momentum

Angular momentum (moment of momentum) of a material point rotating relative to an axis, is a value equal to the product of the momentum of point at a distance from the axis of rotation:

$$L_i = p_i r_i = m_i \upsilon_i r_i. \tag{5.27}$$

Since $\upsilon_i = \omega r_i$ and $J_i = m_i r_i^2$, then

$$L_i = m_i \omega r_i r_i = m_i r_i^2 \omega = J_i \omega. \tag{5.28}$$

Angular momentum of a body rotating relative to an axis is equal to the sum of angular momentums of the momentums of the points that make up the body:

$$L = \sum_{i=1}^{N} J_i \omega. \tag{5.29}$$

Since the angular velocity of all points of a perfectly rigid body is the same, taking ω outside of the sign of the sum [see (5.29)], we obtain:

$$L = \omega \sum_{i=1}^{N} J_i = \omega J. \tag{5.30}$$

(J is a monomeric link).

$$\overline{L} = \vec{\omega} J. \tag{5.31}$$

Thus angular momentum is equal to the product of the moment of inertia of the point by its angular velocity. It follows that the directions of angular momentum and angular velocity vectors coincide. The unit of angular momentum is *kilogram-meter squared per second* ($\mathrm{kg \cdot m^2 \cdot s^{-1}}$).

It is expedient to compare formula (5.31) with a similar formula for momentum in translational motion.

Kinetic energy of a rotating body

When a body rotates, its total kinetic energy is composed of the kinetic energies of individual points of the body. For a perfectly rigid body:

$$E_\mathrm{K} = \sum_{i=1}^{N} \frac{m_i \upsilon_i^2}{2} = \sum_{i=1}^{N} \frac{m_i \omega^2 r_i^2}{2} = \frac{\omega^2}{2} \sum_{i=1}^{N} m_i r_i^2 = \frac{J\omega^2}{2}, \tag{5.32}$$

It is expedient to compare expression (5.32) with a similar expression for translational motion.

After differentiating (5.32), we obtain the elementary change in the kinetic energy in rotational motion:

$$\mathrm{d}E_\mathrm{K} = J\omega\, \mathrm{d}\omega. \tag{5.33}$$

Basic dynamic equation of rotational motion

Let a perfectly rigid body exposed to external forces, be rotated through a sufficiently small angle $d\alpha$. Let us equate the elementary work of all external forces for such rotation (see (5.13)) to elementary change of kinetic energ (see (5.33)): $M\,d\alpha = J\omega\,d\omega$, whence:

$$M\frac{d\alpha}{dt} = J\omega\frac{d\omega}{dt}.$$

Taking in account (5.2), we cancel this equality by ω:

$$M = J\frac{d\omega}{dt}, \qquad (5.34)$$

whence

$$\varepsilon = M/J, \qquad (5.35)$$

or in the vector form:

$$\vec{\varepsilon} = \frac{\overline{M}}{J}. \qquad (5.36)$$

This is the **basic *dynamics equation of rotational motion*.** From (5.35) it can be seen that the moment of inertia characterizes inertial properties of a body in rotational motion: under the impact of external forces, the greater angular acceleration of a body, the smaller its moment of inertia is.

The basic dynamics equation of rotational motion is as significant as Newton's second law for translational motion. Physical quantities included in this equation are analogous to force, mass, and acceleration, respectively.

From (5.34), it follows that:

$$M = \frac{d(J\omega)}{dt} = J\omega\frac{dL}{dt}. \qquad (5.37)$$

The derivative of the angular momentum of a body by time is equal to the resultant moment of all external forces.

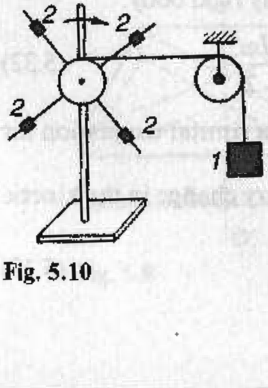


Fig. 5.10

The dependence of angular acceleration on the moment of force and moment of inertia can be shown using the device shown in fig. 5.10. Under the impact of weight *1*, suspended by a thread over a block, the cross is rotating with acceleration. Moving weights *2* to different distances from the axis of rotation, you can change the moment of inertia of the cross. Changing weights, i.e., the moments of force, and moment of inertia, we can see that angular acceleration increases when you increase the moment of force or decrease the moment of inertia.

§ 5.3. LAW OF CONSERVATION OF ANGULAR MOMENTUM

Let us consider the special case of rotational motion, when the total moment of external forces is zero. As it can be seen from (5.37), $dL/dt = 0$ when $M = 0$, from which

$$L = \text{const}, \quad J\omega = \text{const}. \tag{5.38}$$

This provision is known as *the law of angular momentum conservation: if the total moment of all external forces acting on the body is zero, the angular momentum of this body remains constant.*

Omitting the proof, we note that the law of angular momentum conservation is valid not only for a perfectly rigid body.

The most interesting applications of this law are related to rotation of a system of bodies around a common axis. In this case, it is necessary to take into account the vector nature of angular momentum and angular velocities. Thus, for a system consisting of N bodies rotating around the common axis, the law of angular momentum conservation can be thus:

$$L = \sum_{i=1}^{N} J_i\omega_i = \text{const}. \tag{5.39}$$

Let us consider some examples illustrating this law.

A gymnast performing a somersault (fig. 5.11), in the initial phase bends his knees and presses them to his chest, thereby reducing the moment of inertia and increasing the angular velocity of rotation about his horizontal axis passing through the center of mass. At the end of the jump, he straightens his body, the momentum of inertia increases, angular velocity decreases. A figure skater, performing rotation about his vertical axis (fig. 5.12), at the beginning of rotation brings his arms to the body, thereby reducing the moment of inertia and increasing angular velocity. At the end of rotation, a reverse process occurs: the moment of inertia increases and angular velocity decreases as his arms are

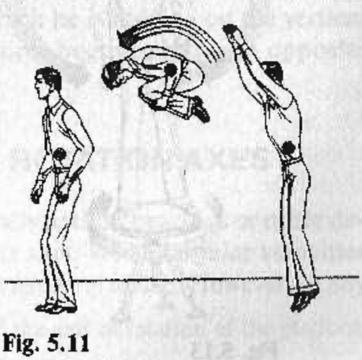


Fig. 5.11

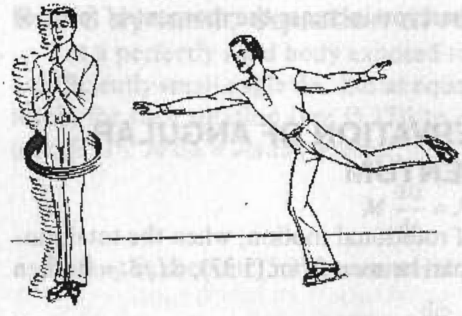


Fig. 5.12

spreading apart, making it easy to stop.

The same phenomenon can be demonstrated on the Zhukovsky bench, which is a light horizontal platform rotating with low friction about its vertical axis. When the position of the subject's arms changes, the moment of inertia and angular velocity changes too (fig. 5.13), angular momentum remains constant. To increase the visual effect, let us provide the subject with dumb-bells. On the Zhukovsky bench, it is possible to demonstrate the vector nature of the law of angular momentum conservation.

The experimenter, standing on the fixed bench, is given a bike wheel rotating around its vertical axis (fig. 5.14, left). In this case, the angular momentum of the person–platform–wheel system is determined only by the angular momentum of the wheel:

$$L = J_P \cdot 0 + J_W\omega_W = J_W\omega_W, \qquad (5.40)$$

here, J_P is the moment of inertia of the person and platform; J_W and ω_W are the moment of inertia and angular velocity of the wheel. As the moment of external forces relative to the vertical axis is zero, then L remains the same ($L = \text{const}$).

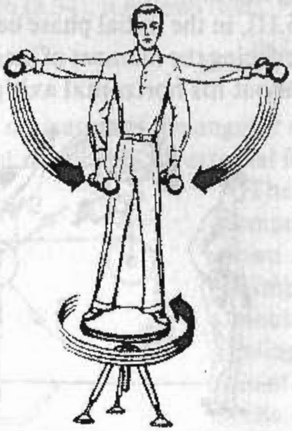


Fig. 5.13

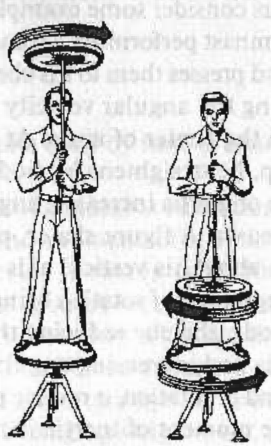


Fig. 5.14

If the experimenter rotates the wheel axis through 180° (fig. 5.14, right), the angular momentum of the wheel will be directed opposite to the original direction and equal to $J_W\omega_W$. Since the angular momentum vector of the wheel will change, and the angular momentum of the system remains the same, the angular momentum of the person and the platform should inevitably change, it will not yet be equal to zero[1]. The total angular momentum of the system in this case is

$$L = J_P\,\omega_P + (-J_W\omega_W) = J_P\,\omega_P + J_W\omega_W. \tag{5.41}$$

The law of conservation of angular momentum makes it possible to equate expressions (5.40) and (5.41):

$$J_W\,\omega_W = J_P\omega_P - J_W\omega_W,$$

or in the scalar form

$$J_W\,\omega_W = J_P\,\omega_P - J_W\,\omega_W, \quad 2J_W\omega_W = J_P\,\omega_P,$$

whence

$$J_P = 2J_W\,\omega_P/\omega_P. \tag{5.42}$$

Formula (5.42) makes it possible to estimate approximately the moment of inertia of the human body together with the platform, for that, it is necessary to measure ω_W, ω_P and to find J_W. The way of measuring angular velocities in case of uniform rotation is known to the reader. Knowing the mass of the wheel and assuming that the mass is mainly distributed along the thin rim, formula (5.22) makes it possible to determine J_W. To reduce the error, you can increase the weight of the bike wheel tin rim, putting special tires on it. The person should be symmetrical relative to the rotation axis.

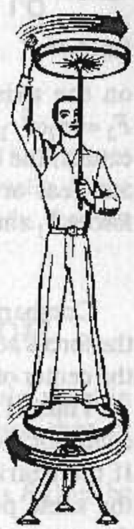


Fig. 5.15

A simpler version of this demonstration is the person himself standing on the Zhukovsky and rotating the wheel, which he is holding on the vertical axis. In this case, the person and the platform begin to move in opposite directions from the wheel (fig. 5.15).

§ 5.4. CONCEPT OF FREE ROTATION AXES

A body rotating around a fixed axis generally acts on bearings or other devices that maintain the fixed position of this axis. When angular velocities and moments of inertia are high, these effects can be sizeable. However, in any

---

[1] A slight mismatch between the wheel axis and the axis of rotation of the platform can be neglected.

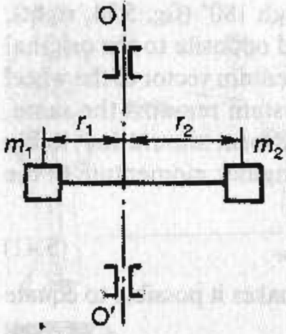


Fig. 5.16

body, it is possible to select such axes, whose direction will be maintained during rotation without any special devices. To understand what condition the choice of such axes should meet, let us consider the following example.

Let a system consisting of two material points with masses m_1 and m_2 and perfectly rigid rod, whose mass is negligible, rotate about axis OO′, fixed in the bearings (fig. 5.16; r_1 and r_2 are the distances from the axis of rotation to the corresponding material points).

Contradirectional centrifugal forces influence on the axis of rotation and, hence, on the bearings: $F_1 = m_1\omega^2 r_1$ and $F_2 = m_2\omega^2 r_2$, where ω is the angular rotation velocity. If these forces do not cancel, the bearings are under constant impact, which can lead to their wear and tear or even destruction. At a certain ratio of masses and distances, forces F_1 and F_2 can be equal, i.e., $m_1\omega^2 r_1 = m_2\omega^2 r_2$, or

$$m_1 r_1 = m_2 r_2. \tag{5.43}$$

Comparing (5.43) with the coordinates of the center of mass, we note that the forces acting on the axis are balanced if the axis of rotation passes through the center of mass.

Thus, if the axis of rotation passes perpendicular to the rod through the center of mass, there will be no impact on this axis from the rotating body. If the bearings are removed, the rotation axis will start to move keeping the same position in space, and the body will continue to move around the same axis.

Axes of rotation, which keep their direction in space without special fastening are called free ones. Examples of such axes are the axis of rotation of the Earth and a spinner, the axis of any thrown and freely rotating body, etc.

An arbitrarily shaped body always has at least three perpendicular axes passing, through the center of mass, which can be free axes of rotation. These axes are called the principal axes of inertia. Although all the three principal axes of inertia are free, the most stable rotation will be one about the axis with the highest momentum of inertia. The thing is that because of the inevitable operation of external forces, such as friction, and since it is difficult to set rotation around a certain axis exactly, rotation about other free axes is unstable.

In some cases, when a body rotates around a free axis with the small moment of inertia, the former replaces this axis with the axis with the highest moment of inertia.

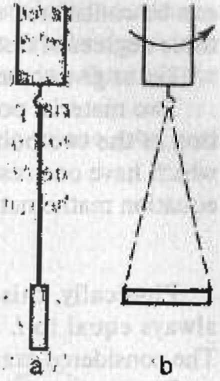


Fig. 5.17

This phenomenon is demonstrated by the following experiment. A cylindrical stick is suspended to the electric motor by a thread, which can move around its geometric axis (fig. 5.17, a). The moment of inertia relative to this axis is $J_1 = mR^2/2$. If angular velocity is high enough, the stick will change its position (fig. 5.17, b). The moment of inertia relative to the new axis is $J_2 = m^2/12$. If $l^2 > 6R^2$, then also $J_2 > J_1$. Rotation about the new axis will be stable.

In practice, the reader can independently verify that rotation of a matchbox thrown is stable relative to the axis passing perpendicular to its greater face, and is unstable or less stable relative to the axes passing perpendicular to other faces (see fig. 5.8).

Rotation of animals and humans in free flight and in different jumps goes on around their free axes with the highest or lowest moment of inertia. Since the position of the center of mass depends on the posture of the body, then for different ones there will be different free axes too.

§ 5.5. CONCEPT OF DEGREES OF FREEDOM

Three independent coordinates set the position of a free material point in space: x, y, z. If the point is not free, but moves, e.g., on some surface, then all three coordinates cannot be be independent.

Let the motion of a material point go along a sphere with radius R given by the equation:

$$x^2 + y^2 + z^2 = R^2.$$

If x and y are considered to be independent, then

$$z = \pm\sqrt{R^2 - x^2 - y^2} \tag{5.44}$$

For instance, let us assume $x = 2$, $y = 3$, $R = 6$, then[1] $z = \pm\sqrt{23}$. So, in this example, only two of the three coordinates are independent variables.

Independent variables characterizing the position of a mechanical system are called degrees of freedom.

A free material point has three degrees of freedom, in the foregoing example there are two degrees of freedom. Since a molecule of the monoatomic gas can be considered a material point, therefore, such a free molecule also has three degrees of freedom.

---

[1] If an imaginary value is obtained for the dependent coordinate from (5.44), it means that the independent coordinates chosen do not correspond to any points located on the sphere of the radius given.

Let us give some other examples.

Two material points 1 and 2 are rigidly connected to each other. The position of the two points is determined by six coordinates x_1, y_1, z_1, x_2, y_2, z_2, which have one restriction and one connection, expressed in the form of an equation mathematically:

$$(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 = l^2.$$

Physically, this means that the distance between the material points is always equal to l. In this case, the number of degrees of freedom is five. The considered example is a model of a diatomic molecule.

Three material points 1, 2 and 3 are rigidly connected with each other. Nine coordinates characterize the position of such a system: x_1, y_1, z_1, x_2, y_2, z_2, x_3, y_3, z_3. However, three connections between the points determine the independence of only six coordinates. The system has six degrees of freedom. Since the position of three points that do not lie on the same straight line determines the position of a rigid body uniquely, a perfectly rigid body also has six degrees of freedom.

Triatomic and polyatomic molecules have the same number of degrees of freedom (six), if these molecules are considered as rigid formations.

In real polyatomic molecules, atoms are in vibrational motion, so the number of degrees of freedom of such molecules is more than six.

The number of degrees of freedom determines not only the number of independent variables characterizing the position of the mechanical system, but also, which is a very significant fact, the number of independent displacements of the system. Thus, three degrees of freedom of a free material point mean that any motion of the point can be broken down to independent displacements along three coordinate axes. Since the point does not have dimensions, it makes no sense to talk about its rotation. Thus, the material point has three degrees of freedom of translational motion. A material point on a plane, sphere, or another surface has two degrees of freedom of translational motion. A motion of a material point along a curve line (a conditional example is a train moving along the rails) corresponds to one degree of freedom of translational motion.

A perfectly rigid body rotating around a fixed axis has one degree of freedom of rotational motion. A wheel of a train has two degrees of freedom: the first one for rotational motion and the other one for translational motion (motion of the wheel axis along the rail). Six degrees of freedom for a perfectly rigid body means that any motion of that body can be broken down to the components: the motion of the center of mass is breaks down into three translational motions along the coordinate axes, and rotation breaks down into three simpler rotations relative to the coordinate axes passing through the center of mass.

In fig. 5.18–5.20, there are hinged joints corresponding to one, two, and three degrees of freedom.

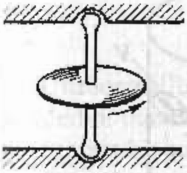


Fig. 5.18

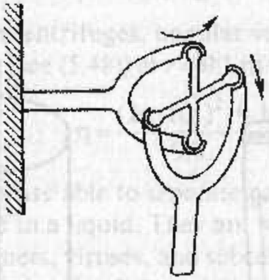


Fig. 5.19

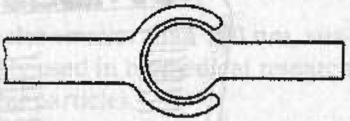


Fig. 5.20

§ 5.6. CENTRIFUGATION

Centrifugation is separation of heterogeneous systems, such as particles from liquids, where the former are determined by the system rotation.

Let us consider separation of inhomogeneous systems in the gravity field. Let us assume that there is aqueous suspension of particles with different density. In the course of time, thanks to action of gravity and buoyancy force F_A the particles become stratified: particles with density higher than water density sink; particles with density lower than water density rise to the surface. The resulting force acting, for instance, on a denser individual particle is equal to:

$$F_p = mg - F_A = \rho_1 Vg - \rho Vg = (\rho_1 - \rho) Vg,$$

where ρ_1 is the density of particle material; ρ is the water density; V is the volume of particle.

If there is little difference between the values of ρ_1 and ρ, then force F_p is small and stratification (sedimentation) goes on rather slowly. In a centrifuge (separator), such separation is coercive when the media rotate.

Let us consider the physical aspect of this phenomenon.

Let the working volume of the centrifuge (fig. 5.21: a is outward appearance; b is the diagram of the working volume) be fully filled with homogeneous liquid. Let us select mentally a small volume V of this liquid located at distance r from the rotation axis OO′. For uniform rotation of the centrifuge, the centripetal force acts on the selected volume in addition to gravity and buoyancy force that balance each other. This is the force from the liquid surrounding the volume. It is surely directed to the axis of rotation and is equal to:

$$F = m\omega^2 r = \rho V \omega^2 r, \tag{5.45}$$

where ρ is the liquid density.

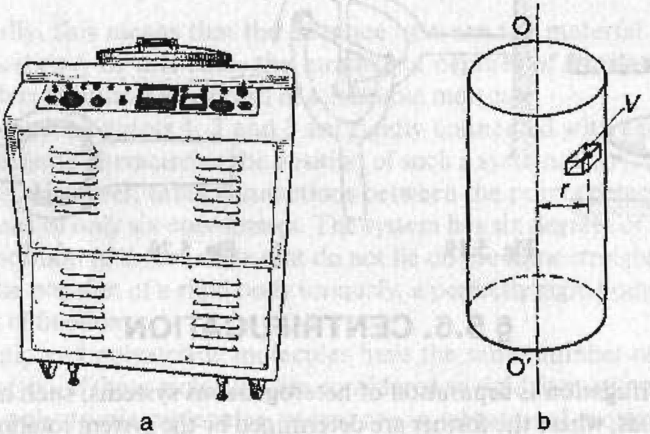


Fig. 5.21

Let us assume now that the separated volume V is a particle to be separated, the density of its material is ρ_1 ($\rho_1 \neq \rho$). The force acting onto the particle from the surrounding liquid will not change as you can see from formula (5.45).

In order for the particle to move together with the liquid, the centripetal force acting onto the particle should be equal to:

$$F_1 = m_1\omega^2 r = \rho_1 V \omega^2 r, \tag{5.46}$$

where m_1 is the particle mass, and ρ_1 is the corresponding density.

If $F > F_1$, then the particle moves to the axis of rotation. If $F < F_1$, then the impact on the particle from the liquid will not be enough to keep it on its circular trajectory, and the particle will begin to move under its own momentum to the periphery. The separation effect is determined by the difference of force F, acting onto the selected particle from the liquid, and the value of centripetal force F_1, which causes its motion along the circumference:

$$F_{\text{cf}} = (F - F_1) = m_1\omega^2 r = (\rho - \rho_1) V \omega_2 r. \tag{5.47}$$

This expression shows that the centrifugation effect is higher the more the difference of the densities of the separated particles and the liquid, and depends on the angular velocity of rotation to a great extent[1].

Let us compare separation by centrifugation with separation by gravity:

$$\eta = \frac{F_{cf}}{F_p} = \frac{(\rho_1 - \rho)V\omega^2 r}{(\rho_1 - \rho)Vg} = \frac{\omega^2 r}{g}. \qquad (5.48)$$

In modern ultracentrifuges, angular velocity can be as high as $\omega = 2\pi \times$ $\times\, 10^3$ rad/s, whence [see (5.48)] at $r = 0.1$ m we have:

$$\eta = \frac{(2\pi \cdot 10^3)^2 \cdot 0.1}{9.8} \approx 4 \cdot 10^5.$$

Ultracentrifuges are able to separate particles smaller than 100 nm, suspended or dissolved in a liquid. They are widely used in biomedical research to separate biopolymers, viruses, and subcellular particles.

The velocity of separation is especially important in biological and biophysical studies, as the state of objects under study can change to a great extent the course of time.

---

[1] Gravity and buoyancy forces were not taken into account in the development of formula (5.47) since they are directed along the axis of rotation and do not have a fundamental effect on centrifugation.

Chapter 6

Some problems of biomechanics

Biomechanics is a section of biophysics, which deals with mechanical properties of living tissues and organs, as well as mechanical phenomena going on both the entire body and its individual organs. In short, biomechanics is mechanics of living systems.

§ 6.1. JOINTS AND LEVERS IN A HUMAN MUSCULOSKELETAL SYSTEM

Moving parts of mechanisms are usually connected to other moving or fixed parts. A movable connection of several links forms kinematic connection. The human body is an example of kinematic connection.

Let us consider a system consisting of two members *A* and *B*, connected by axis OO (fig. 6.1). It is a uniaxial two-member connection. When member *B* is support, member *A* has one degree of freedom as a body rotating around a support axis.

Examples of uniaxial articulation in the human body are the humeroulnar, ankle, and phalanx joints. They make permit only flexion and extension with one degree of freedom. Let us increase the two-member system by one member with axis O′O′, which is parallel to axis OO (fig. 6.2). With support member *C*, all points of member *B* have one degree of freedom, including axis OO, which can move along a circumference. Member *A*, rotating around axis OO, has one more degree of freedom.

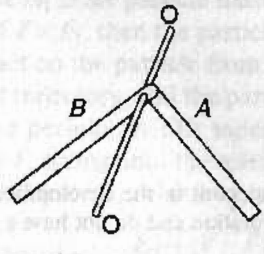


Fig. 6.1

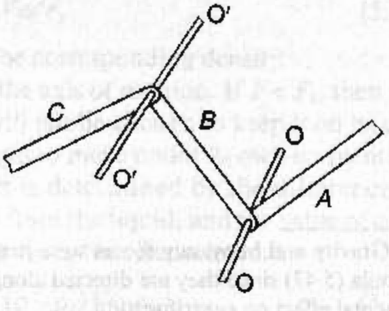


Fig. 6.2

Thus, in a uniaxial three-member system[1], the support member has no freedom of movement, the second member has one degree of freedom, and the third one has two degrees of freedom. The phalanges of the fingers are connected by joints to be uniaxial joints. The nail phalanx has two degrees of freedom relative to the main phalanx and one degree of freedom relative to the middle phalanx.

A biaxial joint permits rotation of its members around two mutually perpendicular axes (see fig. 5.19). It has two degrees of freedom of rotation. Such biaxial connection is carried out in the human body by two close joints: atlantooccipital and atlanto-axial joint. The first joint has a horizontal axis directed from the right shoulder to the left one. It implements cranial rotation forward and back. The axis, which is the cervical vertebra adjacent to atlas, has a small cylindrical spine, which together with the atls ring forms a uniaxial cylindrical joint with the vertical axis. This joint ensures rotation of the head around its vertical axis.

The three-axis joint provides rotation around three mutually perpendicular axes. An example of such a joint is given in fig. 5.20 (ball and socket joint). This joint has three degrees of freedom of rotation. It can seen in the human hip joint. The articular cavity of the pelvis has a shape of an almost perfect hemisphere. The head of femur entering the cavity has the corresponding shape (fig. 6.3).

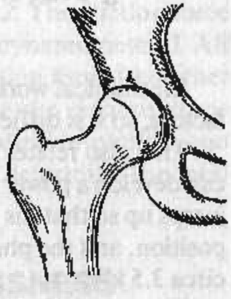


Fig. 6.3

Addition of new members increases kinematic mobility. Thus, the skull, due to some mobility of vertebral joints (although rather limited) has all six degrees of freedom.

The human musculoskeletal system, consisting of articulated bones of the skeleton and muscles, is, from the viewpoint of physics, a totality of levers held by a human in equilibrium.

In anatomy, they distinguish force levers, in which there is a gain in force but a loss in displacement, and velocity levers in which there is a loss of force, but a gain in the velocity of motion. A good example of a velocity lever is the lower jaw. The acting force is generated by the masticatory muscle. The counter force — resistance of the crushed food — acts on the teeth. The arm of the acting force is much shorter than that of the counter forces, so the masticatory muscle is short and strong. When you need to chew something solid, a person uses molars, and the arm of the force of resistance decreases.

If we consider the skeleton as a set of individual members, making up the integral body, it will turn out, that in a normal standing position, all these members form a system in extremely unstable equilibrium. For instance, the body support is represented by spherical surfaces of the hip joint. The center of mass of the trunk is located above the support, which iwith a spherical bearing creates unstable equilibrium. The same applies to the knee joint, and to the ankle joint too. All these members are in a state of unstable equilibrium.

---

[1] The concept of "uniaxial system" does not characterize the number of axes, which can be more than one, but the same direction of all axes.

In a normal standing position the center of mass is located exactly on the same vertical line with the centers of the hip, knee and ankle joints, 2–2.5 cm below the sacropromontory and 4–5 cm above the hip axis. Thus, this is the most unstable state of the members of the skeleton. Moreover, if the whole system is in equilibrium, it is only due to continuous tension of the supporting system of muscles.

§ 6.2. MECHANICAL WORK OF A HUMAN. ERGOMETRY

Mechanical work that a human is able to do during a day depends on many factors, so it is difficult to name any maximum value.

This also relates to power. For instance, with short-term efforts, a human can develop a power circa several kilowatts. If an athlete with the mass of 70 kg jumps up so that his center of mass rises by 1 m in relation to his or her normal position, and the phase of jump-off lasts 0.2 s, then he or she develops a power circa 3.5 kW.

When walking, a human performs work, as the energy is expended for periodic small lifting of the body and acceleration and deceleration of the limbs, mainly the legs.

The work for changing the kinetic energy of the limbs can be calculated using formula (5.32). A human weighing 75 kg walking with the velocity of 5 km/h develops a power of circa 60 W. When his or her velocity increases, this power increases rapidly, attaining 200 W for the velocity of 7 km/h. When riding a bike, the position of the center of mass of the human body changes much less than in walking and the acceleration of the legs is also less. Therefore, the power consumed when riding a bike is much less: 30 W when the velocity is 9 km/h and 120 W when the velocity is 18 km/h.

The work is zero if there is no displacement. Therefore, when a weight is on a support or stand, or is suspended by a thread, the gravity does not do any work. However, each of us knows fatigue of the muscles of his or her arm and shoulder if you keep a dumb-bell motionless in your outstretched arm. Similarly, the back and lumbar region muscles get tired if we put a load on the back of a sitting human. In both cases, the load is motionless and no work is done. The fatigue testifies that the muscles do work. Such work is called ***static muscle work***.

Actually, there is no statics (motionlessness) as it is understood in mechanics. There are very small and frequent, imperceptible to the eye contractions and relaxations, and at the same time, the work against gravity. Thus, static work of a human is actually usual dynamic work. To measure the work of a human, devices called ***ergometers*** are used. The corresponding section of measurement technology is called ***ergometry***.

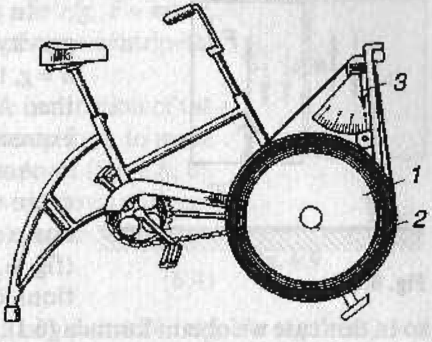


Fig. 6.4

An example of an ergometer is a bike with a brake (bike ergometer; fig. 6.4). Over the rim of the rotating wheel *1*, there is steel band *2*. The friction force between the band and the wheel rim is measured with dynamometer *3*. All work of a tested human is spent for overcoming the friction force (the other types of work are neglected). Multiplying the length of the rim circumference by the friction force, we find the work done during each wheel revolution, and knowing the number of rotations and the test time, we determine the total work and the average power.

§ 6.3. G-FORCE AND WEIGHTLESSNESS

Under normal conditions, people are under the impact of gravity and normal force. In absence of acceleration, these forces are equal and oppositely directed. This state is natural for a person.

Upon accelerated motion of a system, special conditions, called g-forces and weightlessness, can occur.

Let us consider some examples.

Let the human be in an elevator (in a rocket), which is rising with acceleration a (fig. 6.5). The gravity force mg and normal force N act onto the human. According to Newton's second law,

$$\vec{N} + m\vec{g} = m\vec{a}$$

or in the scalar form, taking into account directions of the forces, we obtain

$$N - mg = ma, \quad N = m(g + a). \qquad (6.1)$$

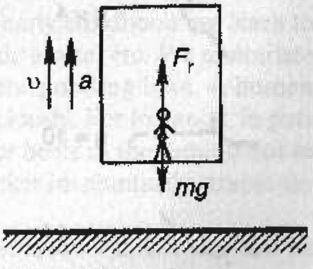


Fig. 6.5

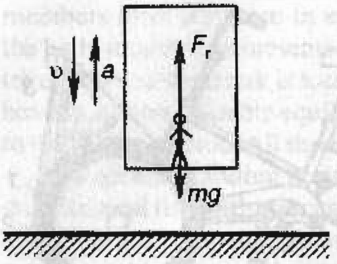


Fig. 6.6

In this case, normal force is more than gravity ($N > mg$), and g-forces occur. Thus, if $a = g$, then $N = 2mg$ (double g-force); if $a = 2g$, then $N = 3mg$ (triple g-force), etc. G-force is expressed by the formula $\eta = N/(mg)$.

Another example: a person is in an elevato cabin (inside a space capsule) that is moving down with deceleration, i.e., falling down (fig. 6.6). Directions of forces and accelerations correspond to the previous example so in this case we obtain formula (6.1). The human is experiencing g-forces.

G-forces can exert a significant influence upon the human body becaus in these states, there is blood outflow, mutual pressure of the viscera changes, their deformation occurs, etc. Therefore, a human is able to withstand only limited g-forces. Fig. 6.7 sketchily shows body positions and corresponding values of g-forces, which a healthy human body can withstand for at least a few minutes without any severe disorders.

In space medicine, large centrifuges are used to train people for g-forc as well as in similar experiments on animals. In such systems (fig. 6.8, a) two supports can be conventionally imagined: horizontal one with the forc $N_1 = mg$, and vertical support N_2, which imparts the body centripetal acceleration and is equal to $m\omega^2 r$. The resultant of these two forces is directed at angle α to the horizontal plane and is equal to:

$$F = \sqrt{(mg)^2 + (m\omega^2 r^2)^2} = m\sqrt{g^2 + \omega^4 r^2}, \quad (6.2)$$

at that $\tan\alpha = mg/(m\omega^2 r) = g/(\omega^2 r)$.

In this case, g-force is determined by the formula:

$$\eta = \frac{F}{mg} = \frac{m\sqrt{g^2 + \omega^4 r^2}}{mg} = \sqrt{1 + \frac{\omega^4 r^2}{g^2}}. \quad (6.3)$$

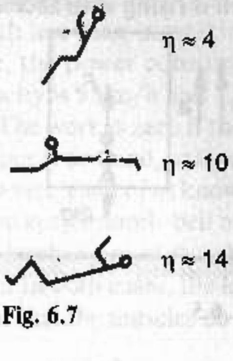


Fig. 6.7

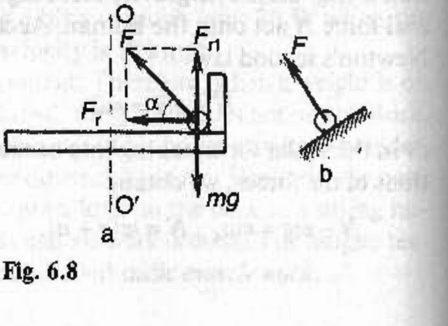


Fig. 6.8

When $\omega^2 r \gg g$ from (6.3) we have $\eta \approx \omega^2 r/g$, $F \approx m\omega^2 r$, and $\tan\alpha \approx 0$. Normal force is mainly acting as centripetal force.

It is practically possible, changing the inclination of the chair, in which the test subject is in the centrifuge, to make force F be always perpendicular to the support (fig. 6.8, b).

If an elevator (or a spacecraft) is moving down with acceleration (fig. 6.9) or upward with deceleration, then

$$mg - N = ma \text{ or } N = m(g - \alpha). \qquad (6.4)$$

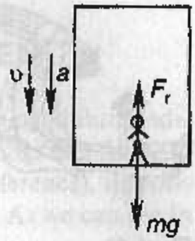


Fig. 6.9

As you can see, normal force is less than gravity: $N < mg$. If $a = g$ then $N = 0$ (condition of weightlessness). This is a state when external forces acting on a system do not cause mutual pressure of the particles of the system onto one another.

For biological objects, weightlessness is an unusual condition, although in everyday life there are short periods of partial weightlessness: jumping, swings, beginning of downward motion of an express elevator, etc.

Absence of normal force in weightlessness generates general deconditioning and the related decrease in efficiency and in muscle mass; the bone tissue undergoes demineralization. Therefore, in weightlessness conditions, astronauts have to do special physical exercises or wear special suits, which, complicating movements, make it possible to give additional work to the muscles.

Under normal conditions, in the top part of the body, the hydrostatic pressure of the blood (ρgh) is less than in the lower part. In weightlessness, the blood is evenly distributed through the body; this means that the upper part of the body is overflown with blood compared with the normal state, the head feels heavy, there face is swollen.

The vestibular apparatus (see § 6.4) will react to weightlessness as if there is no gravitational field, and vestibular disorders will occur.

Let us consider in more detail the specifics of the human body motion in conditions of weightlessness.

People learn the laws of mechanics from their early childhood: we learn to sit, stand, go, run, do physical exercises, work, ride a bike, etc. We assimilate all this without theoretical knowledge of the corresponding laws. A human gets used to doing mechanical operations unconsciously. For instance, in putting the shot, a human instinctively digs his or her heels in the ground not to fall from the recoil; striking with a hammer, a worker involuntarily strains the muscles, preventing rotation of the body, etc.

It is a paradox, but a human gets so used to the laws of mechanics that he begins to notice their manifestation only in special, rare, and unusual cases.

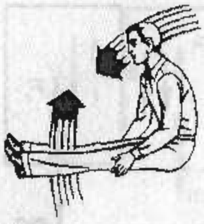


Fig. 6.10

Such specifics and practically important manifestations of the laws of mechanics include human motor activity in conditions of weightlessness, or, as they say, in unsupported space. It is not difficult to calculate, using the law of conservation of momentum, that if a human with a mass 100 kg in weightlessness throws his body with the mass of 0.1 kg with a velocity of 3 m/s, then he starts to move in the opposite direction with a velocity of 0.3 cm/s. If a throw is done with an arm swing, the human body will begin to rotate. This is an unusual manifestation of the laws of conservation of momentum and angular momentum in comparison to terrestrial conditions. A human can stop only by interacting with other bodies. If in the state of weightlessness a human wants to execute half-lever position, which is quite simply done by gymnasts under normal conditions, the motion of his or her legs will cause, in accordance with the law of conservation of angular momentum, counter-rotation of the body (fig. 6.10). Rotation of the body in weightlessness, including free fall, is done by rotating the limbs. Thus, tapered rotational movements of the arm above the head will cause the body to move around the axis of symmetry (fig. 6.11).

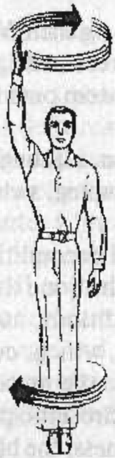


Fig. 6.11

If a human screws a nut in weightlessness, he will start to move in the opposite direction.

In the conditions of weightlessness, the same well-known Newton's laws are valid, but because of the unusual conditions, a human has to get used to moving in weightlessness. Sudden movements with head, arms or legs, throwing any objects can change the motion of the human body to a great extent.

This is taken into account by astronauts both for preparing for space flights and during the flight. The first human of the planet that went into outer space, A.A. Leonov, writes in his book that "...after some training, even in unsupported "floating" in weightlessness, the human can quickly and accurately align his body at any direction solely by his or her muscle efforts, without any technical facilities." Further: "Apparently, in weightlessness, in presence of the smallest point of rest, it is possible to do any work without noticeable disorders of motor coordination!"[1].

---

[1] *Леонов А.А., Лебедев В.И.* Психологические особенности деятельности космонавтов. М., 1971. С. 215, 217. (*A.A. Leonov, V.I. Lebedev.* Psychological Features of Astronaut's Activities. Moscow, 1971. pp. 215, 217).

§ 6.4. VESTIBULAR APPARATUS AS INERTIAL ORIENTATION SYSTEM

Under normal conditions, the position of a free-hanging pendulum indicates the direction of gravity (fig. 6.12, a). If a pendulum belongs to a frame of reference moving with acceleration (non-inertial frame of reference), its position depends on the acceleration of the system (fig. 6.12, b). As we can see in the figure, according to Newton's second law:

$$\vec{F}_e + m\vec{g} = \vec{F}_r = m\vec{a},$$

where the resulting force is equal to

$$F_r = mg\tan\alpha, \text{ or } ma = mg\tan\alpha,$$

whence

$$a = g\tan\alpha. \qquad (6.5)$$

Therefore, even a simple mathematical pendulum can theoretically be used to determine the absolute value and direction of the system acceleration.

A more convenient indicator of the system acceleration is the device shown in fig. 6.13, the body of known mass is fixed by six springs. Deformation of the springs can determine the value and direction of the force acting on the body, and, hence, the acceleration of the system, taking into account the acceleration of gravity. Such indicators are used in inertial navigation, which was developed due to solving space researxh issues.

In fact, if you know the acceleration of a system, such as a rocket, at any given time you can find the dependence of velocity on time:

$$\upsilon = \int a\,dt. \qquad (6.6)$$

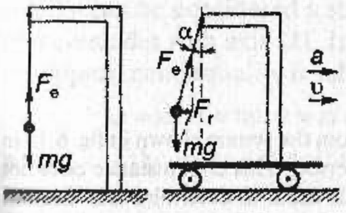


Fig. 6.12

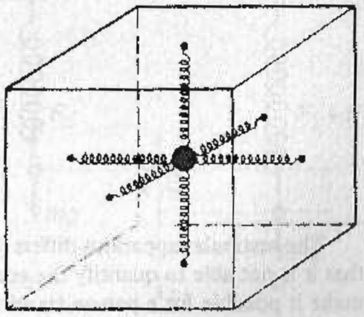


Fig. 6.13

Having found $\upsilon = f(t)$, it is possible to determine the location of the system at any moment of time:

$$x = \int \upsilon_x \mathrm{d}t,\ y = \int \upsilon_y \mathrm{d}t,\ z = \int \upsilon_z \mathrm{d}t. \qquad (6.7)$$

Thus, it is possible to determine independently its location, velocity, and acceleration at any time without the help of means outside the rocket.

The corresponding devices are called ***inertial orientation systems***.

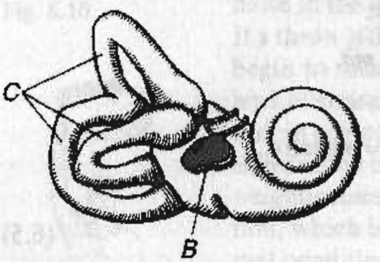


Fig. 6.14

In the human body there is an organ, which is also, in essence, an inertial orientation system: the vestibular apparatus[1]. It is located in the inner ear and consists of three perpendicular semicircular canals C and cavity, which is the vestibule B (fig. 6.14). On the inner surface of vestibular walls and in a part of semicircular canals, there are groups of sensitive nerve cells with free endings in the form of hairs. Inside the vestibule and semicircular channels, there is a gelatinous mass (endolymph), containing small crystals of calcium phosphate and carbonate (otoliths). Accelerated motion of the head a causes change of endolymph and otoliths, which is perceived by nerve cells (through the hairs). The vestibular apparatus, like any other physical system, does not distinguish gravity from effects occurring from accelerated motion of the system.

Our body has adapted to the action of gravity; the cells of the vestibular apparatus that correspond to habitual information transmit it to the brain, so states of both weightlessness and g-force are perceived by us through the vestibular apparatus (and other organs) as unusual states to which it has to get adapted. If a person is exposed to periodical impacts on their vetsibular system, like in ship motions, this can lead to a special state called ***seasickness***.

---

[1] The vestibular apparatus differs fundamentally from the system shown in fig. 6.13 in that it is not able to quantify the acceleration of a person. This circumstance does not make it possible for a person traveling in a closed car cabin to determine the location of the car.

Chapter 7
Mechanical oscillations and waves

Repetitive motions or changes of state are called oscillations (alternating electric current, motion of the pendulum, cardiac performance, etc.). All oscillations, irrespective of their nature, follow some general laws. Oscillations propagate in a medium in the form of waves. Mechanical oscillations and waves are considered in this chapter.

§ 7.1. HARMONIC OSCILLATIONS

Among various types of oscillations, the simplest form is a ***harmonic oscillation***, i.e., one in which the oscillating quantity varies depending on time according to the law of sines and cosines.

Suppose that a material point of mass m is suspended in the spring (fig. 7.1, a). In this position, elastic force F_1 balances gravity mg. If you pull the spring to the distance x (fig. 7.1, b), then a greater elastic force will act on the material point. The change in spring force, according to Hooke's law, is proportional to the change in the length of spring or to displacement x of the point:

$$F = -kx, \tag{7.1}$$

where k is spring stiffness; the minus sign indicates that the force is always directed towards equilibrium: $F<0$ when $x>0$, $F>0$ when $x<0$.

Another example:

A mathematical pendulum is deviated from the equilibrium position by a small angle α (fig. 7.2). Then the pendulum trajectory can be considered a straight line that coincides with axis OX. In this case, the approximate equality is valid:

$$\alpha \approx \sin\alpha \approx \tan\alpha \approx x/l,$$

where x is the material point displacement relative to the equilibrium position; l is the length of pendulum thread.

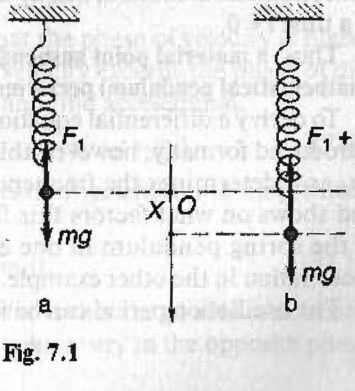


Fig. 7.1

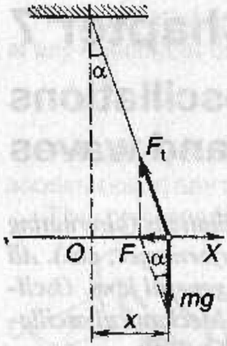


Fig. 7.2

The forces acting onto the material point (see fig. 7.2) are force F_t of thread tension and gravity mg. Their resultant is equal to:

$$F = -mg\tan\alpha = -mgx/l = -kx, \qquad (7.2)$$

where

$$k = mg/l. \qquad (7.3)$$

Comparing (7.2) and (7.1), we see that in this example the resultant force is like spring force, because it is proportional to displacement of the material point and directed towards equilibrium. ***Such forces, non-elastic in their nature, but similar in properties to the forces occurring in small deformations of elastic bodies, are called quasi-elastic.***

Substituting (7.2) into Newton's second law formula, we obtain $-kx = m(d^2x/dt^2)$. Replacing

$$\omega_0^2 = \frac{k}{m}, \qquad (7.4)$$

we have the second order differential equation:

$$\frac{d^2x}{dt^2} = -\omega_0^2 x. \qquad (7.5)$$

Its solution generates the harmonic law:

$$M = A\cos(\omega_0 t + \varphi_0), \qquad (7.6)$$

where $\omega_0 t + \varphi_0$ is the phase of oscillations; φ_0 is the initial phase (when $t = 0$); ω_0 is circular frequency; A is their amplitude.

The amplitude and initial phase of oscillations are determined by the initial conditions of motion, i.e., by the position and velocity of the material point at a time $t = 0$.

Thus, **a material point suspended by a spring (spring pendulum) or a thread (mathematical pendulum) performs harmonic oscillations.**

To derivy a differential equation of harmonic oscillation (7.6), value ω_0 was introduced formally: however, this is justified from the point of view of physics, as it determines the frequency of oscillations of the system $\nu = \omega_0/(2\pi)$ and shows on what factors this frequency depends: the elasticity and mass of the spring pendulum in one example, the length of thread and free fall acceleration in the other example.

The oscillation period can be found from the formula:

$$T = 2\pi/\omega_0. \qquad (7.7)$$

Using (7.4), we obtain the ***spring pendulum period***:

$$T = 2\pi\sqrt{m/k}; \tag{7.8}$$

replacing k with expression (6.3), *we obtain the period of mathematical* pendulum:

$$T = 2\pi\sqrt{l/g}. \tag{7.9}$$

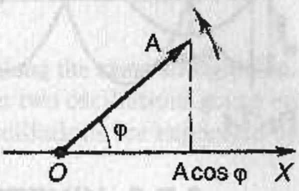


Fig. 7.3

It is very convenient to depict harmonic oscillations with the help of vector diagrams. This method is as follows. From the origin of the abscissa axis, we draw vector $\overline{A}$, whose projection onto axis OX is equal to $A\cos\varphi$ (fig. 7.3) If vector A rotates at angular velocity ω_0 counterclockwise uniformly, then $\omega = \omega_0 t + \omega_0$, where ω is the initial value of φ, and the projection of vector $\overline{A}$ onto axis OX changes in the course of time according to the law (7.6). In this presentation, the oscillation amplitude is equal to the absolute value of uniformly rotating vector $\overline{A}$, the initial phase of oscillations is equal to the angle between vector $\overline{A}$ and axis OX, the initial phase is equal to the initial angle, circular frequency of oscillations is the angular velocity of rotation of vector $\overline{A}$, displacement of the oscillating point is a projection of vector $\overline{A}$ onto axis OX.

To find the velocity of a material point in a harmonic oscillation, you need to take the derivative of expression (7.6) with respect to time:

$$\upsilon = dx/dt = -A\omega_0 \sin(\omega_0 t + \varphi_0) = -\upsilon_{max} \sin(\omega_0 t + \varphi_0), \tag{7.10}$$

where $\upsilon_{max} = A\omega_0$ is the maximum velocity (velocity amplitude).

On the basis of trigonometric formulas we transform (7.10):

$$\upsilon = \upsilon_{max} \cos[(\pi/2) + (\omega_0 t + \varphi_0)]. \tag{7.11}$$

Comparing (7.11) and (7.6), we notice that the phase of velocity is higher by $\pi/2$ than the phase of displacement, i.e., velocity exceeds the phase of displacement by $\pi/2$. Let us differentiate (7.10) and find acceleration:

$$a = d\upsilon/dt = -A\omega_0^2 \cos(\omega_0 t + \varphi_0) = -a_{max} \cos(\omega_0 t + \varphi_0), \tag{7.12}$$

where $a_{max} = A\omega_0^2$ is the maximum acceleration (acceleration amplitude). Instead of (7.12) we write down:

$$a = a_{max} \cos[\pi + (\omega_0 t + \varphi_0)]. \tag{7.13}$$

From comparison of (7.13) and (7.6) it follows that the phases of acceleration and displacement differ by π, i.e., these values vary in the opposite phase.

Graphical time dependences of displacement, velocity and acceleration are shown in fig. 7.4, their vector diagrams are shown in fig. 7.5.

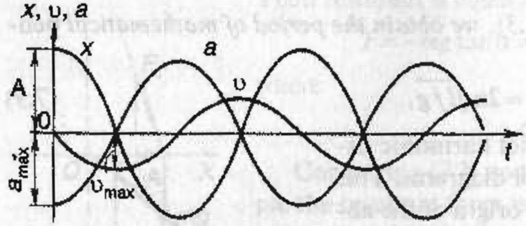


Fig. 7.4

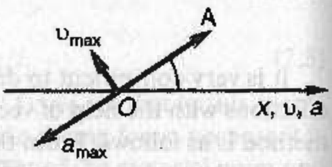


Fig. 7.5

§ 7.2. KINETIC AND POTENTIAL ENERGIES OF OSCILLATORY MOTION

The kinetic energy of an oscillating material point can be calculated by the known formula using the expression (7.10):

$$E_K = {}^1\!/_2 m υ_{max}^2 \sin^2(ω_0 t + φ_0) = {}^1\!/_2 mA^2 ω_0^2 \sin^2(ω_0 t + φ_0) = \\ = {}^1\!/_2 kA^2 \sin^2(ω_0 t + φ_0). \qquad (7.14)$$

The potential energy of oscillating motion is found on the basis of the general formula for the potential energy of elastic deformation $E_P = {}^1\!/_2 kx^2$ and using the expression (7.6):

$$E_P = {}^1\!/_2 kA^2 \cos^2(ω_0 t + φ_0). \qquad (7.15)$$

By adding up kinetic (7.14) and potential (7.15) energy, we obtain the total mechanical energy of the oscillating material point:

$$E = E_K + E_P = {}^1\!/_2 kA^2 \sin^2(ω_0 t + φ_0) + {}^1\!/_2 kA^2 \cos^2(ω_0 t + φ_0) = \\ = {}^1\!/_2 kA^2[\sin^2(ω_0 t + φ_0) + \cos^2(ω_0 t + φ_0)] = {}^1\!/_2 kA^2. \qquad (7.16)$$

In absence of friction forces, the total mechanical energy of the system does not change:

$$E = {}^1\!/_2 kA^2 = {}^1\!/_2 m ω_0^2 A^2. \qquad (7.17)$$

Graphically, the dependence of kinetic, potential and total mechanical energy of an oscillating system on time is shown in fig. 7.6.

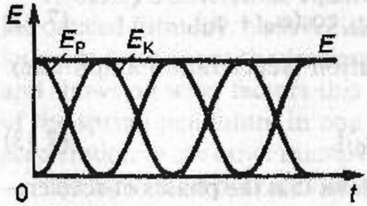


Fig. 7.6

§ 7.3. SUPERPOSED HARMONIC OSCILLATIONS

A material point can simultaneously be involved in several oscillations. In this case, to find the equation and trajectory of the resultant motion, the oscillations should be superposed. The simplest case is when we superpose harmonic oscillations.

Let us consider two such problems.

Superposition of harmonic oscillations directed along the same straight line. Let the material point simultaneously participate in two oscillations going on along the same straight line. Analytically, these oscillations are expressed by the following equations:

$$x_1 = A_1 \cos(\omega_{01} t + \varphi_{01}), \quad x_2 = A_2 \cos(\omega_{02} t + \varphi_{02}).$$

Let us assume that the frequencies of oscillations to be superposed are the same ($\omega_{01} = \omega_{02} = \omega_0$), then the resulting displacement of the point is:

$$x_1 = A_1 \cos(\omega_{01} t + \varphi_{01}), \quad x_2 = A_2 \cos(\omega_{02} t + \varphi_{02}). \tag{7.18}$$

Let us do the superposition using a vector diagram. Let us draw the position of vectors $\overline{A_1}$ and $\overline{A_2}$ at the initial point of time (fig. 7.7), the angles between these vectors and axis *OX* are equal to initial phases of summed oscillations φ_{01} and φ_{02}. Vector $\overline{A}$ is the amplitude of the resulting oscillation. Since $\overline{A_1}$ and $\overline{A_2}$ are rotating with the same angular velocity, their sum is vector $\overline{A}$ moving with the same angular velocity, i.e., the resulting motion is harmonic with circular frequency ω_0:

$$x = A\cos(\omega_0 t + \varphi_0). \tag{7.19}$$

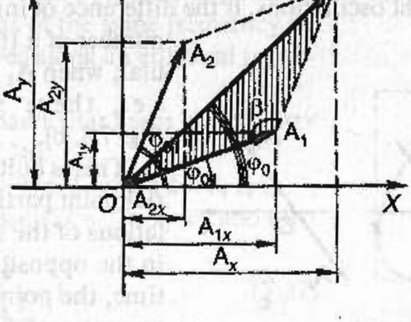


Fig. 7.7

Let us express amplitude A of this oscillation and the initial phase φ_0 through specified values A_1, A_2, φ_{01} and φ_{02}.

Applying the cosine theorem to the triangle shaded in fig. 7.7, we obtain:

$$A^2 = A_1^2 + A_2^2 - 2A_1A_2\cos\beta.$$

Since $\cos\beta = -\cos(\pi - (\varphi_{02} - \varphi_{01})] = \cos(\varphi_{02} - \varphi_{01})$, then

$$A = \sqrt{A_1^2 + A_2^2 + 2A_1A_2\cos(\varphi_{02} - \varphi_{01})}. \tag{7.20}$$

As it is shown in fig. 7.7, $\tan\varphi_0$ is equal to the ratio of projection of A onto axis *Y* to the projection of A onto axis *X*, i.e., A_y/A_x. Taking into account that the projection of the sum is equal to the sum of projections, we have:

$$A_y = A_{1y} + A_{2y} = A_1\sin\varphi_{01} + A_2\sin\varphi_{02},$$

$$A_x = A_{1x} + A_{2x} = A_1\cos\varphi_{01} + A_2\cos\varphi_{02},$$

$$\tan\varphi_0 = A_y/A_x = (A_1\sin\varphi_{01} + A_2\sin\varphi_{02})/(A_1\cos\varphi_{01} + A_2\cos\varphi_{02}). \tag{7.21}$$

Thus, the problem is solved: using formulas (7.20) and (7.21) we can find the amplitude and initial phase of the resulting oscillation. Expression (7.20) yields the following particular cases:

1) $\varphi_{02} - \varphi_{01} = 2k\pi$, $\cos 2k\pi = +1$, where $k = 0, 1, 2, \ldots$, and then

$$A = \sqrt{A_1^2 + A_2^2 + 2A_1A_2\cos(\varphi_{02} - \varphi_{01})}. \tag{7.22}$$

i.e., the amplitude of the resulting oscillation is equal to the sum of amplitudes of constituent oscillations, if the difference of initial phases is an even number of π (fig. 7.8, a);

2) $\varphi_{02} - \varphi_{01} = (2k+1)\pi$, $\cos(2k+1)\pi = -1$, then

$$A = \sqrt{A_1^2 + A_2^2 - 2A_1A_2} = |A_1 - A_2|, \tag{7.23}$$

i.e., the amplitude of the resulting oscillation is equal to the difference of amplitudes of constituent oscillations, if the difference of initial phases is an odd number of π (fig. 7.8, b). In particular, when $A_1 = A_2$ we have $A = 0$, i.e., there is no oscillation (fig. 7.8, b).

This is quite obvious: if a material point participates in two oscillations of the same amplitude and in the opposite phase at the same time, the point is fixed. If the frequencies of summed oscillatio

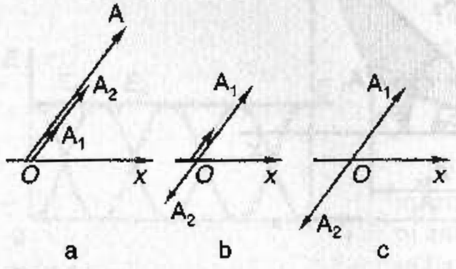


Fig. 7.8

are not the same, the complex oscillation will no longer be harmonic.

It is of interest when frequencies of superposed oscillations differ from each other little: $\omega_{01} \approx \omega_{02}$.

The resulting oscillation is similar to harmonic, but with a slowly varying amplitude (amplitude modulation). Such oscillations are called ***beats*** (fig. 7.9).

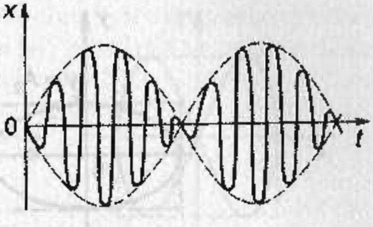


Fig. 7.9

Superposition of mutually perpendicular harmonic oscillations. Let the material point participate in two oscillations simultaneously: one oscillation is directed along axis *OX*, the other one — along axis *OY*. The oscillations are described by the following equations:

$$x = A_1\cos(\omega_{01}t + \varphi_{01}); \quad y = A_2\cos(\omega_{02}t + \varphi_{02}). \qquad (7.24)$$

Let us assume that oscillation frequencies are the same, i.e.,

$$x = A_1\cos(\omega_0 t + \varphi_{01}); \quad y = A_2\cos(\omega_0 t + \varphi_{02}). \qquad (7.25)$$

Equations (7.25) determine the trajectory of the material point in parametric form. If in these equations different values of *t* are substituted, it is possible to determine coordinates *x* and *y*, and the set of points with the corresponding coordinates is the trajectory.

More clearly, the trajectory can be represented as a dependence $y = f(x)$, to obtain which one should exclude time from the equations (7.25). After making mathematical transformations, we obtain equation of an ellipse:

$$\frac{x^2}{A_1^2} + \frac{y^2}{A_1^2} - 2\frac{xy}{A_1A_2}\cos(\varphi_{02} - \varphi_{01}) = \sin^2(\varphi_{02} - \varphi_{01}). \qquad (7.26)$$

Thus, when a material point participates in two mutually perpendicular harmonic oscillations of the same frequency simultaneously, it moves along an elliptical trajectory (fig. 7.10).

There are a few particular cases as corollary facts of expression (7.26):

1) $\varphi_{02} - \varphi_{01} = (2k + 1)\pi/2$, where $k = 0, 1, 2, \ldots$, $\cos[(2k + 1)\pi/2] = 0$, $\sin[(2k + 1)\pi/2] = 1$ and then

$$\frac{x^2}{A_1^2} + \frac{y^2}{A_2^2} = 1. \qquad (7.27)$$

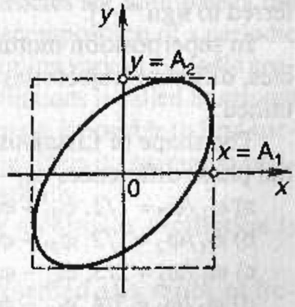


Fig. 7.10

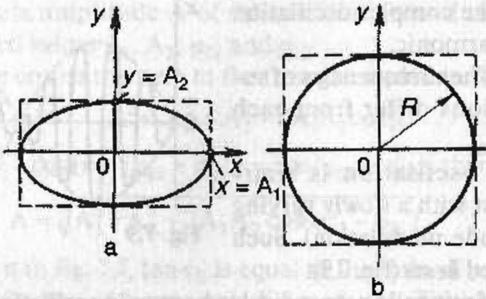


Fig. 7.11

This is the canonical form of an ellipse equation corresponding to its symmetrical position relative to the coordinate axes (fig. 7.11, a). From (7.27), when $A_1 = A_2 = R$ (fig. 7.11, b) we get equation of a circumference with radius R:

$$x^2 + y^2 = R^2; \tag{7.28}$$

2) $\varphi_{02} - \varphi_{01} = 2k\pi$, where $k = 0, 1, 2, \ldots$, $\cos k\pi = \pm 1$, $\sin^2 k\pi = 0$, and then

$$\frac{x^2}{A_1^2} + \frac{y^2}{A_2^2} \pm 2\frac{xy}{A_1 A_2} = 0 \tag{7.29}$$

or after transformations

$$\left(\frac{x}{A_1} \pm \frac{y}{A_2}\right)^2 = 0, \quad \frac{x}{A_1} \pm \frac{y}{A_2} = 0, \quad y = \pm\frac{A_2}{A_1}x. \tag{7.30}$$

This is the equation of a ***straight line*** into which the ellipse degenerates [fig. 7.12, and corresponds to sign "+" in the equation (7.30); fig. 7.12, b is referred to sign "–"].

In superposition mutually perpendicular oscillations of different frequencies, different trajectories of a material point called ***Lissajous figures*** are obtained.

The shape of Lissajous figures depends on frequency ratio ω_1/ω_2 and initial phase differences $\varphi_{02} - \varphi_{01}$ of oscillations being composed (fig. 7.13):

a) $\omega_1/\omega_2 = 1/2$, $\varphi_{01} - \varphi_{02} = 0$;
b) $\omega_1/\omega_2 = 1/2$, $\varphi_{01} - \varphi_{02} = \pi/2$;
c) $\omega_1/\omega_2 = 2/3$, $\varphi_{01} - \varphi_{02} = \pi/2$;
d) $\omega_1/\omega_2 = 3/4$, $\varphi_{01} - \varphi_{02} = \pi/2$.

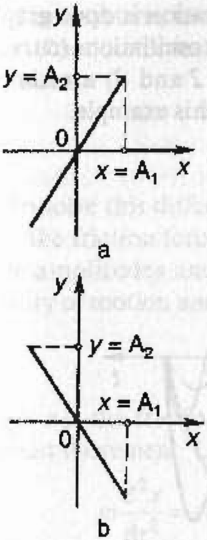


Fig. 7.12

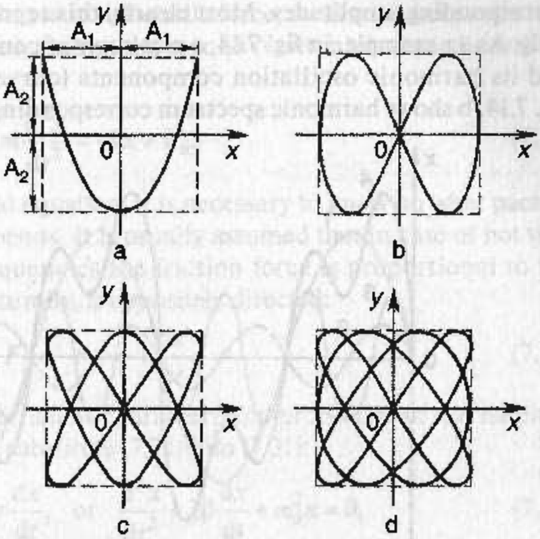


Fig. 7.13

§ 7.4. COMPOUND OSCILLATION. HARMONIC SPECTRUM OF A COMPOUND OSCILLATION

As it can be seen in § 7.3, superposition of oscillations generates more compound forms of oscillations. For practical purposes, the opposite operation can be needed: expansion of compound oscillations as a series of simple, usually harmonic oscillations.

Fourier has shown that a periodic function of any complexity can be represented as a sum of harmonic functions whose frequencies are multiples of the frequency of a compound periodic function. This decomposition of a periodic function into harmonic functions and, hence, representing various periodic processes (mechanical, electrical, etc.) as harmonic oscillations is called harmonic or Fourier. There are mathematical expressions that make it possible to find harmonic function components. Special devices called analyzers do harmonic analysis of oscillations, also for medical purposes, automatically.

The set of harmonic oscillations representing a complex oscillation is called ***harmonic spectrum of a complex oscillation***.

Harmonic spectrum can be conveniently represented as a series of frequencies (or circular frequencies) of individual harmonics together with their corresponding amplitudes. Most clearly, this representation is done graphically. As an example, in fig. 7.14, a plot curve of complex oscillations (curve *4*) and its harmonic oscillation components (curves *1*, *2* and *3*) are shown; fig. 7.14, b shows harmonic spectrum corresponding to this example.

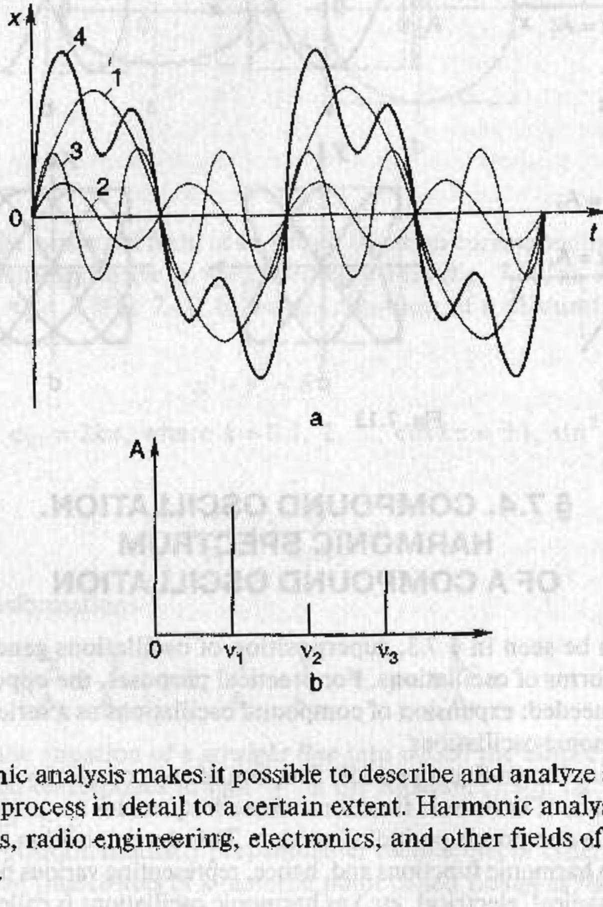


Fig. 7.14

Harmonic analysis makes it possible to describe and analyze any complex oscillatory process in detail to a certain extent. Harmonic analysis is applied in acoustics, radio engineering, electronics, and other fields of science and technology.

§ 7.5. DAMPED OSCILLATIONS

The foregoing consideration of harmonic oscillations does not take into account friction and resistance forces that exist in real systems. The effect of these forces changes the nature of motion to a great extent, and the oscillation becomes *damped*.

If in a system, in addition to quasi-elastic force, resistance forces of a medium (friction forces) are operating, the second Newton's law can be written as follows:

$$m\frac{d^2x}{dt^2} = -kx + F_{fr}. \tag{7.31}$$

To solve this differential equation, it is necessary to know on what parameters the friction force depends. It is usually assumed that in case of not very large amplitudes and frequencies the friction force is proportional to the velocity of motion and, naturally, is oppositely directed:

$$F_{fr} = -r\upsilon = -r\frac{dx}{dt}, \tag{7.32}$$

where r is the friction coefficient characterizing properties of the medium to resist movement. Let us substitute (7.32) into (7.31):

$$m\frac{d^2x}{dt^2} = -kx - r\frac{dx}{dt}, \quad \text{or} \quad \frac{d^2x}{dt^2} + 2\beta\frac{dx}{dt} + \omega_0^2 x = 0, \tag{7.33}$$

where $2\beta = r/m$; $\omega_0^2 = k/m$; β is the attenuation coefficient; ω_0 is the circular frequency of the system proper oscillations.

The solution of (7.33) depends on the sign of difference very much:

$$\omega^2 = \omega_0^2 - \beta^2,$$

where ω is the circular frequency of damped oscillations. When $\omega_0^2 - \beta^2 > 0$, circular frequency ω is a real quantity and the solution of (7.33) will be the following:

$$x = A_0 e^{-\beta t}\cos(\omega t + \varphi_0). \tag{7.34}$$

The plot curve of this function is shown in fig. 7.15 with a solid line *1*; dashed line *2* represents the amplitude change:

$$A = \pm A_0 e^{-\beta t}. \tag{7.35}$$

The period of damped oscillations depends on the friction coefficient and is determined by the formula:

$$T = \frac{2\pi}{\omega} = \frac{2\pi}{\sqrt{\omega_0^2 - \beta^2}}. \tag{7.36}$$

In case of a very low friction ($\beta^2 \ll \omega_0^2$), the period of damped oscillation is close to the period of undamped free oscillation:

$$T \approx 2\pi/\omega_0.$$

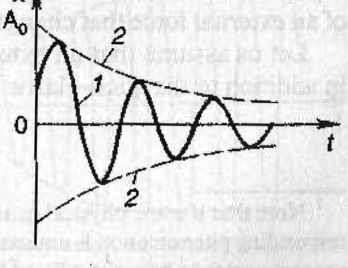


Fig. 7.15

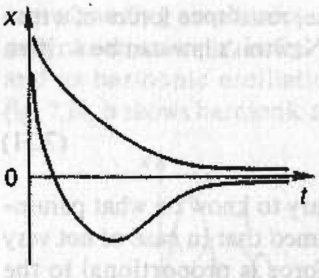


Fig. 7.16

The rate of decrease of oscillation amplitude is determined by the ***damping coefficient***: the higher β, the stronger the inhibiting affect of the medium is, and the quicker the reduction of the amplitude. However, in practice, the degree of damping is often characterized by a ***logarithmic decrement of damping***, meaning a value equal to the natural logarithm of two successive oscillation amplitudes separated by an interval of time equal to the period of oscillation:

$$\lambda = \ln\frac{A(t)}{A(t+T)} = \ln\frac{A_0 e^{-\beta t}}{A_0 e^{-\beta(t+T)}} = \ln e^{\beta t}.$$

hence, the damping coefficient and logarithmic decrement of damping are related by a rather simple dependence:

$$\lambda = \beta T. \tag{7.37}$$

In case of strong damping ($\beta^2 \gg \omega_0^2$), from formula (7.36) it is clear that the oscillation period is an imaginary value. The motion in this case is already called *aperiodic* one[1]. Possible aperiodic motions are presented in the form of plot curves in fig. 7.16. This case is considered in more detail relatively to electric phenomena in Chapter 18.

Undamped (see § 7.1) and damped oscillations are called ***free*** oscillations. They occur because of initial displacement or initial velocity and occur in absence of external influence due to the originally accumulated energy.

§ 7.6. FORCED OSCILLATIONS. RESONANCE

Forced oscillations are oscillations that occur in a system with participation of an external force that changes according to the periodic law.

Let us assume that an external driving force acts onto a material point, in addition to the quasi-elastic force and friction forces:

$$F = F_0 \cos\omega t,$$

---

[1] Note that if some physical quantity takes imaginary values, it means that some corresponding phenomenon is unusual, extraordinary. In this example, its peculiarity is the process ceases to be a periodic one.

where F_0 is amplitude; ω is the circular frequency of the driving force. Let us compose a differential equation (Newton's second law):

$$m\frac{d^2x}{dt^2} = -kx - r\frac{dx}{dt} + F_0 \cos\omega t,$$

or

$$\frac{d^2x}{dt^2} + 2\beta\frac{dx}{dt}\omega_0^2 x = f_0 \cos\omega t, \tag{7.38}$$

where $f_0 = F_0/m$.

The solution of the differential equation (7.38) is the sum of two summands. One of them, corresponding to the equation of damped oscillations (7.34), is important only in setting up the oscillations (see fig. 7.15). In the course of time, it can be neglected. The other summand describes the displacement of the material point in stationary forced oscillations:

$$x = A\cos(\omega t + \varphi_0), \tag{7.39}$$

where

$$A = f_0/\sqrt{(\omega_0^2 - \omega^2)^2 + 4\beta^2\omega^2}, \tag{7.40}$$

$$\tan\alpha\ \varphi_0 = -2\beta\omega/(\omega^2 - \omega_0^2). \tag{7.41}$$

As it can be seen from (7.39), the stationary forced oscillation under the influence of a harmonically changing driving force is also harmonic. The frequency of forced oscillation is equal to the frequency of the driving force. Forced oscillations, whose plot curve is shown in fig. 7.17, shift in phase relative to the driving force.

The amplitude of a forced oscillation (7.40) is directly proportional to the amplitude of the driving force and has a complex dependence on the damping coefficient of the medium and circular frequencies of free and forced oscillations. If ω_0 and β are set for the system, then the amplitude of forced oscillations has the maximum value for a certain frequency of driving force called resonance frequency. The phenomenon itself attaining the maximum amplitude for a set ω_0 and β is called resonance.

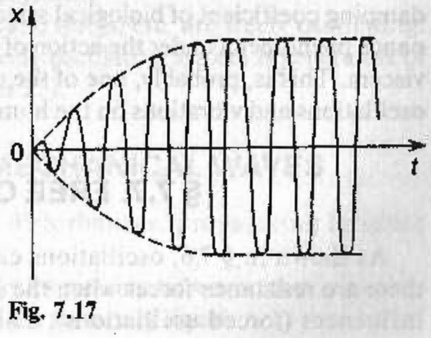


Fig. 7.17

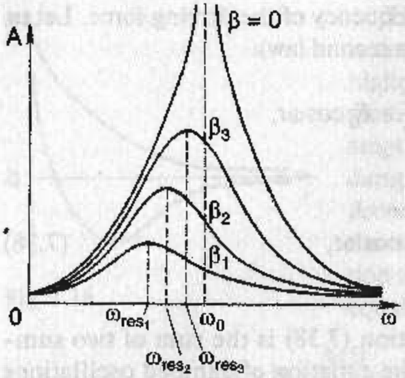


Fig. 7.18

Resonance circular frequency can be found by determining the minimum condition of the denominator in (7.40):

$$\omega_{res} = \sqrt{\omega_0^2 - 2\beta^2}. \qquad (7.42)$$

Substituting (7.42) into (7.40), we find the amplitude for resonance:

$$A_{res} = f_0 / \left(2\beta\sqrt{\omega_0^2 - \beta^2}\right). \qquad (7.43)$$

It can be seen from (7.43) that in absence of resistance ($\beta = 0$) the amplitude of forced oscillations in resonance is infinitely high. At the same time from (7.42) it follows that $\omega_{res} = \omega_0$ is resonance in a system without damping. It happens when the frequency of a driving force coincides with the free frequency of oscillations. Graphic dependence of the amplitude of forced oscillations on the circular frequency of the constraining force at different values of the damping coefficient is shown in fig. 7.18.

Mechanical resonance can be both useful and harmful. The harmful effect of resonance is mainly associated with destruction that it can cause. Thus, in technology, taking in account various vibrations, it is necessary to foresee possible occurrence of resonance conditions, otherwise there can be accidents and disasters. Bodies usually have several natural oscillation frequencies and several resonance frequencies respectively.

If the damping coefficient of human viscera were small, resonance phenomena that would have occurred in these organs under the influence of external vibrations or acoustic waves would have been able to lead to tragic consequences: rupture of organs, lesion of ligaments, etc. However, such phenomena are practically absent when external effects are moderate since the damping coefficient of biological systems is large enough. Nevertheless, resonance phenomena under the action of external mechanical vibrations occur in viscera. This is, probably, one of the causes of negative impact of infrasoun oscillations and vibrations on the human body (see § 8.7 and 8.8).

§ 7.7. FREE OSCILLATIONS

As shown in § 7.6, oscillations can be maintained in the system even if there are resistance forces when the system is periodically exposed externa influences (forced oscillations). This external influence does not depen on the oscillating system itself, while the amplitude and frequency of forced oscillations depend on this external influence.

However, there are also such oscillating systems that regulate periodic recovery of the expended energy themselves and so can oscillate for a long time.

Undamped oscillations existing in any system in the absence of variable external influence are called freely oscillating, and such systems themselves are freely oscillating.

The amplitude and frequency of freely oscillating depend on the properties of a freely oscillating system itself; in contradistinction to forced oscillations, they are not determined by external influences.

In many cases, freely oscillating systems can be represented by three main elements:

1) oscillating system proper;
2) energy source;
3) regulator of energy supply to the oscillating system itself.

An oscillating system (fig. 7.19) affects its regulator through feedback, informing the regulator of the state of the system.

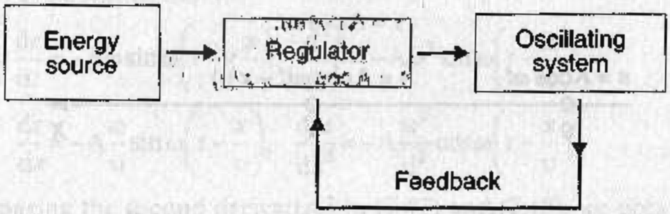


Fig. 7.19

A classic example of a mechanical freely oscillating system is a clock in which the pendulum or balance is an oscillating system, where the spring or raised weight is the source of energy, and the pallet is the regulator of energy supply from the source to the oscillating system.

Many biological systems like the heart, lungs, etc. are freely oscillating. A typical example of electromagnetic freely oscillating system is generators of electromagnetic oscillations (see Chapter 23).

§ 7.8. EQUATION OF MECHANICAL WAVES

A mechanical wave is mechanical disturbances, propagating in space and carrying energy.

There are two main types of mechanical waves: elastic waves, that is spreading of elastic deformations, and waves on the surface of liquid.

Elastic waves occur due to the bonds existing between particles of a medium: motion of one particle from the equilibrium position generates displacement of neighboring particles. This process spreads in space with a finite velocity.

The wave equation expresses the dependence of displacement s of an oscillating point, participating in the wave process, on the coordinate of its equilibrium position and time.

For a wave spreading in a certain direction OX, this dependence is written in a general form:

$$s = f(x, t).$$

If s and x are directed along the same straight line, the wave is ***longitudinal***, if they are mutually perpendicular, then the wave is ***transverse***.

Let us derive a plane wave equation. Let the wave spread along axis X (fig. 7.20) without attenuation so that oscillation amplitudes of all points are the same and equal to A. Let us set the oscillation of the point with the coordinate $x = 0$ (the source of oscillations) with the equation

$$s = A\cos\omega t.$$

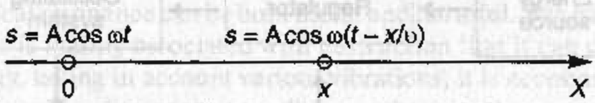


Fig. 7.20

From the origin of the coordinates disturbance will reach a point with some arbitrary coordinate x after time t so the oscillations of this point are late:

$$s = A\cos[\omega(t - r)]. \tag{7.44}$$

Since the time and velocity of the wave propagation are related by the dependence $t = x/\upsilon$, then instead of (7.44) we obtain

$$s = A\cos[\omega(t - x/\upsilon)]. \tag{7.45}$$

This is the ***plane wave equation***, which permits determining the displacement of any point involved in a wave process, at any moment of time. The cosine argument $\varphi = \omega(t - x/\upsilon)$ is called ***phase*** of the wave. The set of points with the same phase at the same time is called ***wave front***. For the considered case, the wave front is plane $x = \text{const}$ (the plane is perpendicular to axis OX), all points of which are corresponded by the same phase at the same time. Whence, the name is ***plane*** wave.

The velocity of propagation of a fixed phase of oscillations is called ***phase*** velocity. Let us assume that $\varphi = \omega(t - x/\upsilon) = \text{const}$. Differentiating this equation, we get $0 = \omega(\mathrm{d}t - \mathrm{d}x/\upsilon)$, from which $\upsilon = \mathrm{d}x/\mathrm{d}t$.

Therefore, the velocity of fixed phase propagation of oscillations is the velocity of wave propagation.

In addition to phase velocity, there is ***group*** velocity, which is introduced when the real wave cannot be represented by one harmonic equation (7.45), but is a composite of a group of sinusoidal waves.

Wavelength is the distance between two points, whose phases at the same time differ by 2π. It is equal to the distance passed by the wave over the period of oscillation:

$$\lambda = T\upsilon. \tag{7.46}$$

The wave equation (7.45) is one of possible solutions of the general partial differential equation describing the process of perturbation propagation in a medium. Such an equation is called wave equation.

To get an idea of wave equation, let us differentiate (7.45) twice by time t and twice by coordinate x:

$$\frac{\mathrm{d}s}{\mathrm{d}t} = -A\omega \sin\omega\left(t - r\frac{x}{\upsilon}\right), \quad \frac{\mathrm{d}^2 s}{\mathrm{d}t^2} = -A\omega^2 \sin\omega\left(t - \frac{x}{\upsilon}\right); \tag{7.47}$$

$$\frac{\mathrm{d}s}{\mathrm{d}x} = -A\frac{\omega}{\upsilon}\sin\omega\left(t - \frac{x}{\upsilon}\right), \quad \frac{\mathrm{d}^2 s}{\mathrm{d}x^2} = -A\frac{\omega^2}{\upsilon^2}\cos\omega\left(t - \frac{x}{\upsilon}\right). \tag{7.48}$$

Comparing the second derivatives in (7.47) and (7.48), we obtain a one-dimensional wave equation:

$$\frac{\mathrm{d}^2 s}{\mathrm{d}x^2} = \frac{1}{\upsilon^2}\frac{\mathrm{d}^2 s}{\mathrm{d}t^2}. \tag{7.49}$$

Solving partial differential equations is beyond the scope of this course. But one solution of (7.45) is known. However, it is important to note the following: if variation of any physical quantity — mechanical, thermal, electrical, magnetic, etc. — meets equation (7.49), it means that the corresponding physical quantity propagates as a wave with velocity υ.

§ 7.9. ENERGY FLOW RATE OF WAVES. UMOV–POYNTING VECTOR

A wave process is associated with transmission of energy. The quantitative characteristic of energy transmitted is an energy flow rate.

Energy flow rate of waves is the ratio of energy transmitted by waves through a certain surface, to the time during which this energy has been transmitted:

$$\Phi = \frac{dE}{dt}.$$

The unit of flow rate of wave energy is *Watt* (W).

Let us establish the relation between the energy flow rate of waves with the energy of oscillating points, and the velocity of wave propagation.

Let us select a volume of the medium in which the wave is propagating in the form of a rectangular parallelepiped (fig. 7.21), whose cross-sectional area is S, and the length of the edge is numerically equal to velocity υ and coincides with the direction of wave propagation. According to this, the energy of oscillating particles will pass within 1 s through area S in the parallelepiped volume equal to $S\upsilon$. This is flow rate of wave energy:

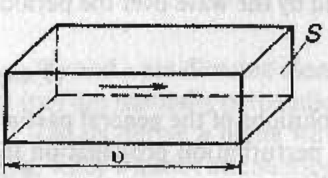


Fig. 7.21

$$\Phi = \omega_p S\upsilon, \tag{7.50}$$

where ω_p is volume density of oscillation motion energy.

The flow rate of wave energy related to an area perpendicular to the direction of wave priopagation is called a density of flow rate of wave energy or wave intensity

$$I = \Phi/S = \omega_p \upsilon,$$

or in the vector form:

$$\bar{I} = \omega_p \bar{\upsilon}. \tag{7.51}$$

The unit of density of flow rate of wave energy is *Watt per square meter* (W/m^2).

Vector I, showing the direction of wave propagation and equal to wave energy flow rate, passing through a unit area, which is perpendicular to this direction, is called ***Umov–Poynting vector***.

Energy transmitted by the elastic wave consists of potential energy of deformation and the kinetic energy of oscillating particles. Let us consider causes on which the volume density of the wave energy depends in this case. If we insert the density of the material ρ instead of the mass of a single particle into (7.17), we obtain

$$\omega_p = rA^2\omega^2/S. \tag{7.52}$$

Substituting (7.52) into (7.51), we have

$$\bar{I} = (\rho A^2\omega^2/2)\bar{\upsilon}. \tag{7.53}$$

Thus, **the Umov–Poynting vector for an elastic wave depends on the density of the medium, the square of oscillation amplitude of particles, the square of oscillation frequency and the velocity of wave propagation.**

§ 7.10. SHOCK WAVES

One common example of a mechanical wave is ***acoustic wave*** (see Chapter 8). In this case, the maximum velocity of oscillations of a single air molecule is a few centimeters per second, even with a sufficiently high intensity, i.e., it is much less than the velocity of the wave (the velocity of sound in air is about 300 m/s). This corresponds, as they say, to small-amplitude perturbations of the medium.

However, in case of grand disturbances (explosion, supersonic motion of bodies, powerful electric discharge, etc.), the velocity of oscillating particles in a medium can become comparable to the velocity of sound, which results in a shock wave.

During an explosion, highly heated products with a large density expand and compress the layers of surrounding air. In the course of time, the volume of compressed air increases. The surface that separates compressed air from free air is called *shock wave* in physics. Sketchily, the density jump of gas, when the shock wave propagates in it, is shown in fig. 7.22, a. For comparison, the same figure shows the change in medium density during the passage of a sound wave (fig. 7.22, b).

The shock wave can have sizeable energy, like in case of a nuclear explosion. Circa 50% of the energy of explosion is expended for formation of a shock wave in the environment. Therefore, the shock wave, reaching biological and technical objects, can cause death, mutilations, and destruction.

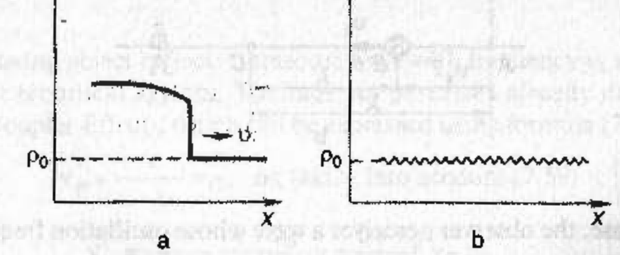


Fig. 7.22

§ 7.11. DOPPLER EFFECT

The Doppler effect is change in the frequency of waves perceived by the observer (wave receiver) due to the relative motion of the wave source and the observer.

Imagine that the observer is approaching with velocity υ_o to a source of waves which is fixed relative to the medium. At this, during the same time interval, he encounters more waves than in absence of motion. It means that perceived frequency ν' is higher than the frequency of the wave emitted by the source. But since the wavelength, frequency and velocity of the wave are related by expression $\nu = \upsilon/\lambda$, then $\nu' = (\upsilon + \upsilon_o)/\lambda$, or taking into account that $\lambda = \upsilon/\nu$:

$$\nu' = \frac{\upsilon + \upsilon_o}{\upsilon}\nu. \tag{7.54}$$

Another case: the source of waves S moves with velocity v to the observer fixed relative to the medium (fig. 7.23, a). Since the source moves after the emitted wave, the wavelength will be smaller than in case of a steady source. In fact; the wavelength is equal to the distance between two points with phase difference 2π. During a time equal to one period, the wave will travel distance λ (fig. 7.23, b), the wave source will travel distance $AB = \upsilon_o T$. At all that, phases of points B and C differ by 2π; hence, the distance between them is equal to wavelength λ' of the wave generated during the motion of the source of emission.

Using fig. 7.23 and knowing that $\nu = \upsilon/\lambda$, let us make some transformations:

$$\lambda' = \lambda - \upsilon_s/\nu = \upsilon/\nu - \upsilon_s/\nu = (\upsilon - \upsilon_s)T/\nu. \tag{7.55}$$

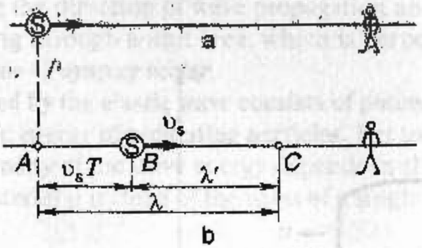


Fig. 7.23

In this case, the observer perceives a wave whose oscillation frequency is

$$\nu'' = \upsilon/\lambda' = [\upsilon/(\upsilon - \upsilon_o)]\nu. \tag{7.56}$$

When the observer and source approach each other, the perceived frequency formula is obtained by substituting ν′ into formula (7.56) [see(7.54)] instead of ν:

$$\nu''' = \frac{\upsilon}{\upsilon - \upsilon_o} \cdot \frac{\upsilon + \upsilon_o}{\upsilon} \nu = \frac{\upsilon + \upsilon_o}{\upsilon - \upsilon_o} \nu. \tag{7.57}$$

As it can be seen from (7.57), when approaching the source of waves and the observer, the perceived frequency is greater than the emitted one. Having changing signs of υ_o and υ_s in (7.57), you can obtain a similar formula for a case when the source moves away from the observer. Thus, it is possible to write a general formula

$$\nu''' = \frac{\upsilon \pm \upsilon_o}{\upsilon \mp \upsilon_o} \nu, \tag{7.58}$$

where the upper signs correspond to approaching the source and receiver of waves, and the lower signs correspond to their moving away from each other.

The Doppler effect is used for determining the velocity of a body in a medium. This is especially important for medical studies. Let us consider in more detail the following case.

Let an ultrasound generator be combined with a receiver in the form of some technical system (fig. 7.24). The technical system is fixed relative to the medium. In the media, the velocity of an object (a body) is υ_0. The generator emits ultrasound with frequency ν_G. As an observer, the moving object perceives frequency ν_1, which can be found by formula (7.54):

$$\nu_1 = \frac{\upsilon + \upsilon_0}{\lambda} = \frac{\upsilon + \upsilon_o}{\upsilon} \nu_G, \tag{7.59}$$

where υ is the velocity of propagation of the mechanical wave (ultrasound).

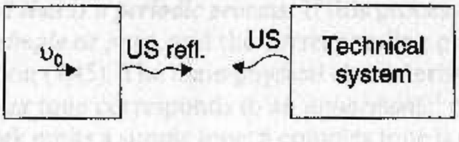


Fig. 7.24

The moving object reflects ultrasonic wave with frequency ν_1 in the direction of the technical system. The receiver perceives already different frequency (Doppler Effect), which can be expressed using formula (7.56):

$$\nu_{re} = \frac{\upsilon}{\upsilon - \upsilon_0} \nu_G, \quad \text{or, taking into account (7.59)}$$

$$\nu_{re} = \frac{\upsilon}{\upsilon - \upsilon_0} \cdot \frac{\upsilon + \upsilon_0}{\nu} \cdot \nu = \frac{\upsilon + \upsilon_0}{\upsilon - \upsilon_0} \nu_G. \tag{7.60}$$

Thus, the frequency difference is

$$\nu_d = \nu_{re} - \nu_G = \frac{\upsilon+\upsilon_0}{\upsilon-\upsilon_0}\nu_G - \nu_G = \frac{\upsilon+\upsilon_0-\upsilon+\upsilon_0}{\upsilon-\upsilon_0} = \frac{2\upsilon_0}{\upsilon-\upsilon_0}\nu_G \qquad (7.61)$$

and called ***Doppler frequency shift***.

In medical applications, the velocity of ultrasound is significantly highe than the velocity of an object ($\upsilon \gg \upsilon_0$). For these cases, from (7.61), we have

$$\nu_d = \frac{2\upsilon_0}{\upsilon}\nu_G. \qquad (7.6$$

The Doppler effect is used to determine the velocity of blood flow (see § 11.5), velocities of valves and walls of the heart (Doppler echocardiography), and other organs.

Chapter 8

Acoustics

Acoustics is a field of physics that studies elastic oscillations and waves from the lowest to highest frequencies ($10^{12}-10^{13}$ Hz). Present-day acoustics covers a wide range of problems; a number of subdivisions are distinguished: physical acoustics, which studies the specifics of elastic waves propagating in different media, physiological acoustics, studying the structure and functions of sound-perceiving and sound-generating organs in humans and animals, etc. In the broad sense of the word, acoustics is defined as the study of sounds, i.e., elastic oscillations and waves in gases, liquids, and solids perceived by a human ear (frequencies from 16 to 20,000 Hz).

§ 8.1. NATURE OF SOUND. PHYSICAL CHARACTERISTICS

Sound oscillations and waves are a particular case of mechanical oscillations and waves. However, due to the importance of acoustic concepts for estimation of auditory sensations, as well as in connection with medical applications, it is expedient to analyze some problems specifically. Conventionally the following sounds are distinguished:

1) tones and musical sounds;
2) noises;
3) sonic booms.

A tone* is a sound *that is a periodic process. If this process is harmonic, then the tone is called *simple* or *pure*, and the corresponding plane sound wave is described by equation (7.45). The main physical characteristic of a pure tone is frequency. A *complex* tone corresponds to an *anharmonic*[1] oscillation. For instance, a tuning-fork emits a simple tone; a complex tone is created by musical instruments, a vocal apparatus (vowels), etc.

A complex tone can be decomposed into simple ones. The lowest frequency ν_0 of such decomposition corresponds to a *fundamental tone*; the other harmonics (*overtones*) have frequencies equal to $2\nu_0$, $3\nu_0$ etc. The set of frequencies with indication of their relative intensity (amplitudes A) are called *acoustic spectrum* (see § 6.4). The spectrum of a complex tone is discrete; fig. 8.1 shows acoustic spectra of the same note ($\nu_0 = 100$ Hz), played on a grand piano (a)

---

[1] Anharmonic is non-harmonic oscillation.

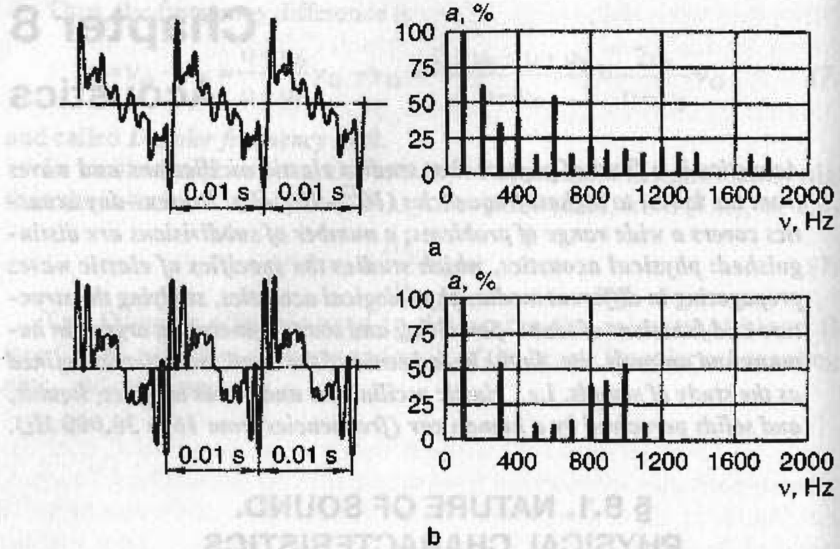


Fig. 8.1

and on a clarinet (b). Thus, the acoustic spectrum is an important physic characteristic of a complex tone.

Noise is a sound characterized by complex nonperiodic time dependence.

A noise is the following sounds: vibration of machinery, applause, nois of a burner flame, rustling, squeaking, consonants, etc.

Noise can be considered as a combination of randomly changing comple tones. If we approximately try to decompose the noise into a spectrum it will turn out that this spectrum will be continuous, like the spectrum obtaine from the combustion noise of Bunsen burner (fig. 8.2).

A sonic shock wave is ***a short-time sound impact: burst, explosion, etc*** *It should not be confused with sonic boom* (see § 7.10).

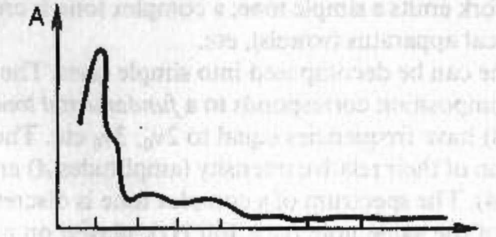


Fig. 8.2

The energy characteristic of sound as a mechanical wave is its intensity (see § 7.9), which can also be expressed by Umov–Poynting vector.

In practice, to estimate a sound it is more convenient to use sound pressure rather than intensity, which additionally occurs due to passage of sound waves in a liquid or gaseous medium. For a plane wave, intensity I is connected with sound pressure ρ by the dependence

$$I = p^2/(2\rho c),^{1}$$

where ρ is the medium density; c is the velocity of sound.

A normal human ear perceives quite a wide range of sound intensities: so, at a frequency of 1 kHz, from $I_0 = 10^{-12}$ W/m^2 or $p_0 = 2 \cdot 10^{-5}$ Pa (hearing threshold) to $I_{\max} = 10$ W/m^2 or $p_{\max} = 60$ Pa (pain threshold). The ratio of these intensities is 10^{13}, so it is more convenient to use logarithmic units (see § 1.1) and logarithmic scale. The scale of sound intensity levels is created as follows: value I_0 is taken as the initial level of the scale, any other intensity I is expressed through the decimal logarithm of its relation to I_0:

$$L_{\text{B}} = \lg(I/I_0), \tag{8.1}$$

and for the sound pressure:

$$L_{\text{B}} = 2\lg(p/p_0).$$

For using decibels, we obtain respectively

$$L_{\text{dB}} = 10\lg(I/I_0) \text{ and } L_{\text{dB}} = 20(p/p_0). \tag{8.2}$$

Sound pressure in gases is measured with a microphone, which consists of a sensor that converts the acoustic value into an electric signal, electronic amplifier and electric gauge (fig. 8.3). This diagram is a more particular case of the general block diagram (see § 21.1).

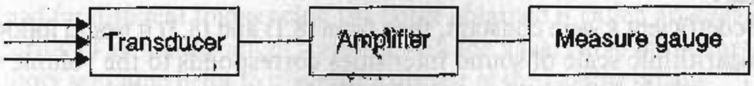


Fig. 8.3

§ 8.2. CHARACTERISTICS OF AUDITORY SENSATION. SOUND MEASUREMENTS

In § 8.1, objective characteristics of sound were considered. They can be estimated by instruments independently from the human ear. However, a sound is an object of auditory sensations, so it is evaluated by a person subjectively.

---

[1] Strictly speaking, in this formula p should be understood as the mean amplitude of sound pressure.

Perceiving tones, a human distinguishes them by their pitch.

A pitch is a subjective characteristic; it is, first, the frequency of the fundamental tone.

To a much lesser extent, the pitch depends on the complexity of the tone and its intensity; a sound of greater intensity is perceived as a sound of a lower tone.

Timbre of the sound is almost solely determined by its spectral composition.

In fig 8.1, different acoustic spectra correspond to different timbres, although the fundamental tone and, therefore, the pitch of the tone are the same.

Volume is a subjective assessment of sound that characterizes the level of auditory sensation.

Despite its subjectivity, volume can be evaluated quantitatively by comparing an auditory sensation from two sources.

The important psycho-physical law of Weber–Fechner is the base of the volume level scale: ***if you increase the stimulus in a geometric progression (i.e., by the same number of times), the sensation increases in an arithmetic progression (i.e., by the same value).***

In case of sound, this means that if the intensity of sound takes a number of consecutive values, e.g. aI_0, a^2I_0, a^3I_0 (a is a coefficient, $a > 1$) and so on, the corresponding sensation of the sound volume is E_0, $2E_0$, $3E_0$, and so on.

This means mathematically that the sound volume is proportional to a logarithm of sound intensity.

If there are two sound stimuli with intensities I and I_0, and if I_0 is the hearing threshold, then, on the basis of the law of Weber–Fechner, their relative volume is associated with the intensities as follows:

$$E = k\lg(I/I_0), \tag{8.3}$$

where k is the coefficient of proportionality depending on frequency and intensity.

If coefficient k were constant, then from (8.1) and (8.3) it would follow that the logarithmic scale of sound intensities corresponds to the volume scale. In this case, the volume of sound, as well as intensity, would be expressed in bels or decibels. However, the strong dependence of k on frequency and intensity of sound does not permit reducing measurement of volume to simply applying formula (8.3).

Conventionally, it is considered that at the frequency of 1 kHz, the volume and intensity scales of sound coincide completely, i.e., $k = 1$ and $E_B = \lg(I/I_0)$, or, by analogy with (8.2):

$$E_{Ph} = 10\lg(I/I_0). \tag{8.4}$$

To distinguish them from the scales of intensity of sound in terms of volume, decibels are called ***phons*** (phon, sing.).

The volume at other frequencies can be measured comparing the investigated sound with the sound at frequency 1 kHz. To do this, using a ***sound generator***[1], a sound with frequency 1 kHz is produced. The intensity of the sound is changed until there is an auditory sensation similar to the sensation of the volume of the sound being studied. The intensity of the sound with frequency of 1 kHz in decibels, measured by a device, is equal to the volume of the sound in phons.

In order to find a match between the volume and intensity of sound for different frequencies, curves of equal volume are used (fig. 8.4). These curves are plotted on the base of average data obtained from people with normal hearing for measurements done by the method described above.

The lower curve corresponds to the intensity of the weakest audible sounds — the threshold of hearing; at all frequencies $E_{Ph} = 0$, at 1 kHz the sound intensity $I_0 = 1$ pW/m^2. The curves above show that the average human ear is most sensitive to frequencies of 2,500–3,000 Hz. Each intermediate curve corresponds to the same volume, but with a different sound intensity for different frequencies. Along an individual curve, corresponding to the volume, you can find intensities, whose frequencies produce the sensation of this volume. Using a family of equal volume curves for different frequencies, one can be find the volume corresponding to a certain intensity. For example, the sound intensity for 100 Hz is 60 dB. What is the volume of that sound? In fig. 8.2, we find the point with coordinates 100 Hz, 60 dB. It lies on the curve corresponding to the volume of 30 phons, which is the answer.

To have some ideas of different nature of the sounds, we give their physical characteristics (Table 8.1).

The method of measuring hearing acuity is called *audiometry*. In audiometry, with a special device (*audiometer*), the threshold of auditory sensation is determined for different frequencies; the curve obtained is called an *audiogram*. Comparing of an audiogram of a patient with the usual curve of the threshold of auditory sensation helps to diagnose a disease of the hearing organs.

A *sound level meter* is *used for objective measuring of noise level.* Its structure corresponds to the diagram shown in fig. 8.3. The characteristics of the sound level meter are close to those of a human ear (see equal volume curves in fig. 8.4), for this purpose corrective electric filters are used for different ranges of volume.

---

[1] A sound generator is an electronic device that generates electrical oscillations with frequencies within the sound range. However, the sound generator itself is not a source of sound. If the oscillation generated by it is sent to a loudspeaker, then there is a sound, whose tone corresponds to the frequency of the generator. The sound generator makes it possible to change the amplitude and frequency of oscillations smoothly.

Table 8.1

Approximate sound nature	Sound intensity, W/m^2	Sound pressure, Pa	Intensity level of sound relative to hearing threshold, db (or level of the sound volume at frequency 1 khz, phons)
Threshold of hearing	10^{-12}	0.00002	0
Cardiac tones heard through a stethoscope	10^{-11}	0.000064	10
Whisper	10^{-10}	0.0002	20
	10^{-9}	0.00064	30
Conversation:			
quiet	10^{-8}	0.002	40
normal	10^{-7}	0.0064	50
loud	10^{-6}	0.02	60
Noise on a busy street	10^{-5}	0.064	80
Shouting	10^{-4}	0.2	80
Noise:			
in a subway train	10^{-3}	0.64	90
of motorcycle (maximum)	10^{-2}	2	100
engine of a plane	10^{-1}	6.4	110
same, but nearby	100	20	120
Threshold of pain	10	64	130

§ 8.3. PHYSICAL FUNDAMENTALS OF SOUND RESEARCH METHODS IN CLINIC PRACTICE

Sound, like light, is a source of information, and this is its main significance.

Sounds of nature, speech of the people around us, noise of operating machines tell us a lot. To imagine the importance of sound for a human, it is enough to deprive yourself temporarily of the ability to perceive sound — simply closing your ears.

Naturally, a sound can be a source of information about the actual state of internal organs. A common sonic method of diagnostics, *auscultation* (listening)

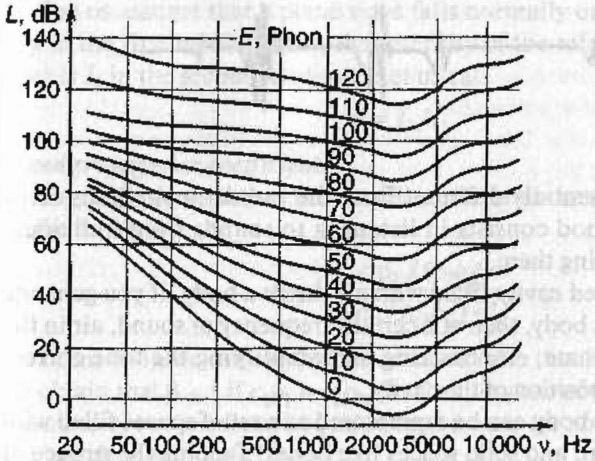


Fig. 8.4

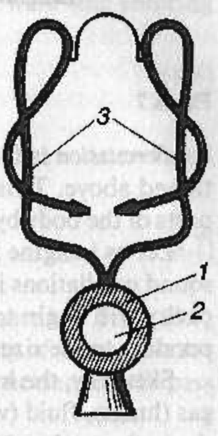


Fig. 8.5

is known since the 2nd century BC. For auscultation, *stethoscope or phonendoscope are used.* A phonendoscope (fig. 8.5) consists of a hollow capsule *1* with a sound-transmitting membrane *2*, which is put to the patient's body; from it, rubber tubes *3* go to the physician's ear. In the hollow capsule, the air column produces resonance, so the sound is amplified and auscultation improves.

In auscultation of the lungs, one listens to breath sounds, different wheezes that are characteristic for diseases. According to changes in cardiac tones and emergence of noises, one can assess the state of cardiac activity. Using auscultation it is possible to determine the presence of peristalsis of the stomach and intestine, listen to fetal heartbeat.

For simultaneous listening to the patient by several researchers for educational purposes or for a consultation, a system that includes a microphone, amplifier and loudspeaker or several phones is used.

To assess the state of cardiac activity, a method similar to auscultation called *phonocardiography* (PCG) is used. This method is graphic recording of cardiac tones and noises and their diagnostic interpretation. Phonocardiogram recording is done by a phonocardiograph (fig. 8.6), consisting of a microphone, amplifier, a system of frequency filters, and recording device. A normal phonocardiogram is shown in fig. 8.7.

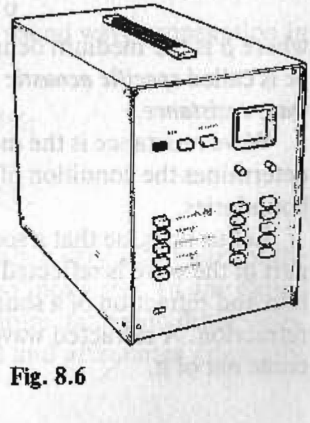


Fig. 8.6

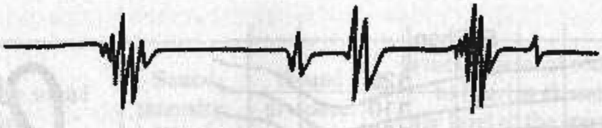


Fig. 8.7

Percussion is fundamentally different from the two sonic methods mentioned above. This method consists in listening to sounds from individual parts of the body by tapping them.

Let us imagine a closed cavity filled with air inside a body. If you generate sound oscillations in this body, then at a certain frequency of sound, air in the cavity will begin to resonate, emphasizing and amplifying the tone corresponding to the size and position of the cavity.

Sketchily, the human body can be represented as a set of spaces filled with gas (lungs), fluid (viscera), and solid spaces like bones. Tapping the surface of the body produces, oscillations with a wide range of frequencies. Within this range, some oscillations will be damped rather quickly, while others, coinciding with proper oscillations of cavities, will be amplified and, due to resonance, will be heard. An experienced physician can assess the state and topography of inner organs according to the tones of percussion sounds.

§ 8.4. WAVE RESISTANCE. SOUND WAVES REFLECTION. REVERBERATION

Sound pressure p depends on velocity υ of oscillating medium particles. Calculations show that

$$\frac{p}{\upsilon}=\rho c \quad \text{or} \quad p=\rho c\upsilon, \tag{8.5}$$

where ρ is the medium density; c is wave velocity in the medium. Production ρc is called ***specific acoustic impedance***, in case of a plane wave, it is also called ***wave resistance***.

Wave resistance is the most important characteristic of the medium, which determines the condition of reflection and refraction of waves at the medium's boundaries.

Let us imagine that a sound wave falls on the boundary of two media. One part of the wave is reflected and the other part is refracted. The laws of reflection and refraction of a sound wave are similar to those of light reflection and refraction. A refracted wave can be absorbed by the second medium, or can come out of it.

Let us assume that a plane wave falls normally on the interface, its intensity in the first medium is I_1, the intensity of the refracted wave (the entering one) is I_2 in the second medium. Let us call

$$\beta = I_2/I_1 \tag{8.6}$$

sound propagation coefficient.

Rayleigh showed that the coefficient of sound propagation is determined by the formula:

$$\beta = 4\frac{c_1\rho_1/(c_2\rho_2)}{[c_1\rho_1/(c_2\rho_2)+1]^2}. \tag{8.7}$$

From (8.6), we can see that the greatest value that β can have is 1. From (8.7) we obtain that $\beta = 1$ if $c_1\rho_1 = c_2\rho_2$. So then, in case of equality of when wave resistances of two media are equal, a sound wave (at normal incidence) will pass the interface without reflection.

If the wave resistance of the second medium is very high compared to the wave resistance of the first medium ($c_2\rho_2 \gg c_1\rho_1$), then, instead of (8.7), we have

$$\beta \approx 4c_1\rho_1/(c_2\rho_2), \tag{8.8}$$

because $c_2\rho_2/(c_1\rho_1) \ll 1$. Let us cite some wave resistances of some materials at 20 °C (Table 8.2).

Table 8.2

Material	Wave resistance, $kg \cdot m^{-2} \cdot s^{-1}$	Material	Wave resistance, $kg \cdot m^{-2} \cdot s^{-1}$
Iron	40,000,000	Rubber	60,000
Concrete	4,800,000	Air	400
Water	1,440,000	Oil	1,350,000

Let us use (8.8) to calculate the coefficient of sound wave propagation in air and water:

$$\beta = \frac{4 \cdot 440}{4\,800\,000}100\% = 0.037\%;$$

$$\beta = \frac{4 \cdot 440}{1\,440\,000}100\% = 0.122\%.$$

These data are impressive: it turns out that only a very small part of the sound wave energy pass from air into concrete and into water. In any closed room, sound is reflected from walls, ceiling, furniture, falls onto other walls, floors, etc.; it is reflected again and is absorbed and attenuates gradually.

So, even after the sound source stops working, there are still sound waves in the room that produce humming. This is especially noticeable in big spacious halls. The process of gradual sound attenuation in closed rooms after switching off the source is called ***reverberation***.

Reverberation, on the one hand, is useful since perception of sound is enhanced at the expense of the energy of the reflected wave, but, on the other hand, excessively long reverberation can impair the perception of speech, music, as each new part of the text overlaps with the previous one. In this connection, usually some optimal reverberation time is stated, which is taken into account for construction of lecture halls, theater and concert halls, etc. Thus, the reverberation time of a packed Column Hall of the House of Unions in Moscow is 1.70 seconds, in a packed Bolshoi Theater it is 1.55 seconds. In these rooms (when they are empty), the reverberation time is 4.55 seconds and 2.06 seconds, respectively.

§ 8.5. PHYSICS OF HEARING

The auditory system connects the direct receiver of the sound wave with the brain.

Using the concepts of cybernetics, we can say that the auditory system receives, processes, and transmits information. Of the entire auditory system, we distinguish the outer, middle and inner ear for consideration of the physical aspect of hearing.

The outer ear consists of auricle *1* and the ear canal *2* (fig. 8.8).

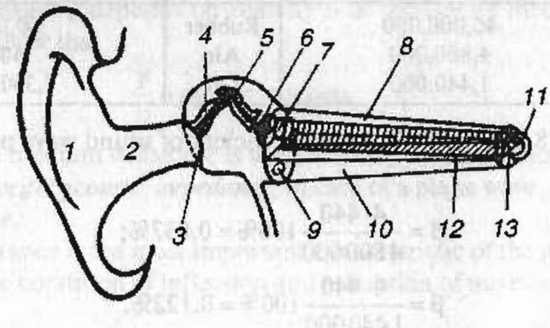


Fig. 8.8

The auricle is not of great importance for human hearing. It helps to determine the localization of the sound source when it is located in the sagittal plane. Let us explain this. A sound from a source gets into the auricle. Depending on the source position in the vertical plane (fig. 8.9), sound waves are differently diffracted on the auricle because of its specific shape. This will lead to different changes in the spectral composition of the sound wave entering the ear canal (in more detail, diffraction problems are discussed in § 24.6). Accumulating experience humans learned to connect change in the spectrum of sound wave with the direction of the sound source (directions *A*, *B* and *C* in fig. 8.9).

Having two sound receivers (ears), humans and animals are able to determine the direction of the sound source in the horizontal plane also (binaural effect; fig. 8.10). This is due to the fact that the sound passes a different distance to each ear, and there is a phase difference in the waves entering the right and left auricle. The relation between these distances (δ) and phase difference ($\Delta\varphi$) is derived in § 24.1 when explaining light interference [see (24.9)]. If the sound source is right in front of the person's face then $\delta = 0$ and $\Delta\varphi = 0$; if the sound source is located on one side in front of an auricle, then into the other auricle, it will come with a delay. We assume that in this case δ is approximately equal to the distance between the auricles. Formula (24.9) makes it possible to calculate the phase difference for $\nu = 1$ kHz and $\delta = 0.15$ m. It is about 180°.

Different directions of a source of sound in the horizontal plane, correspond to phase differences between 0° and 180° (for the values mentioned above). A human with normal hearing is considered to be able to detect a direction of the sound source with the accuracy of 3°, which corresponds to a phase difference of 6°. Therefore, we can assume that a human can distinguish a change in the phase difference of sound waves coming to into his ears, with an accuracy of up to 6°.

In addition to phase difference, the binaural effect is promoted by different sound intensity in the ears, as well as the acoustic shadow from the head for one ear. It is sketchily shown in fig. 8.10 that the sound from the source enters the left ear due to diffraction.

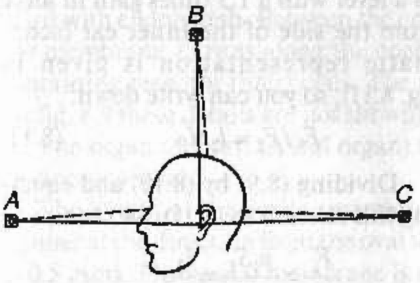


Fig. 8.9

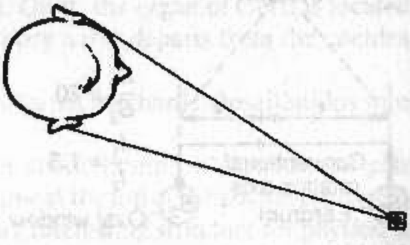


Fig. 8.10

A sound wave passes through the ear canal and is partially reflected from the eardrum *3*. Due to interference from incident and reflected waves, acoustic resonance can occur when the wavelength is four times more than the length of the ear canal. In humans, the latter is approximately 2.3 cm; therefore, acoustic resonance occurs at a frequency of:

$$\upsilon = \frac{c}{\lambda} = \frac{3\cdot 10^2}{4\cdot 2.3\cdot 10^{-2}} = 3\ \textbf{KHz}.$$

The most significant parts of the middle ear are eardrum *3* and auditory ossicles: malleus *4*, incus *5* and stapes *6* with correspondent muscles, tendons and ligaments. The ossicles transmit mechanical oscillations from the air medium of the outer ear to the liquid medium of the inner ear. The liquid medium of the inner ear has a wave resistance approximately equal to the wave resistance of water. As shown above (see § 8.4), only 0.122% of the intensity of the incident sound is transmitted during the direct transition of the sound wave from air to water. This is too little. That is why, the main purpose of the middle ear is to improve transmission of greater sound intensity to the inner ear. In a technical language, we can say that the middle ear matches wave resistances of air and liquid of the inner ear.

At the one end, the system of ossicles is connected by the malleus with the eardrum (area $S_1 = 64\ \text{mm}^2$), at the other end, the stapes connects it with the oval window *7* of the inner ear (area $S_2 = 3\ \text{mm}^2$).

Sound pressure p_1 acts onto the eardrum, which generates the force

$$F_1 = p_1 S_1. \tag{8.9}$$

At the same time, the oval window of inner ear is exposed to force F_2 causing sound pressure p_2 in liquid medium. The relation between them is:

$$F_2 = p_2 S_2. \tag{8.10}$$

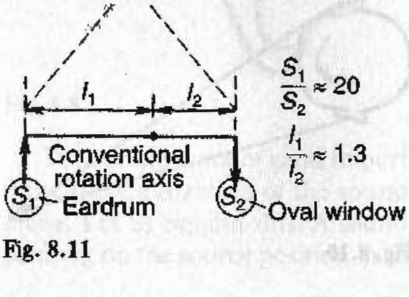

Ossicles system
l_1
l_2
$\frac{S_1}{S_2} \approx 20$
$\frac{l_1}{l_2} \approx 1.3$
Conventional rotation axis
S_1
Eardrum
S_2
Oval window

Fig. 8.11

The system of human ossicles works as a lever with a 1.3 times gain in force from the side of the inner ear (schematic representation is given in fig. 8.11), so you can write down:

$$F_1/F_2 = l_2/l_1. \tag{8.11}$$

Dividing (8.9) by (8.10) and equating this relation to (8.11), we obtain:

$$\frac{F_1}{F_2} = \frac{p_1 S_1}{p_2 S_2} = \frac{l_2}{l_1},$$

whence

$$\frac{p_2}{p_1} = \frac{S_1}{S_2} \cdot \frac{l_1}{l_2} = 20 \cdot 1.3 = 26$$

or in logarithmic units (see § 1.1)

$$L_B = 20\lg(p_2/p_1) = 20\lg 26 = 20 \cdot 1{,}415 \approx 28 \text{ dB}.$$

At this level, the middle ear increases the transmission of external sound pressure to the inner ear.

Another function of the middle ear is attenuating the transmission of oscillations in case of high-intensity sound. This is done by reflex relaxation of the muscles of middle ear ossicles. The middle ear is connected with the atmosphere through the auditory (Eustachian) tube.

The outer and middle ear belong to the sound-conducting system. The sound receiving system is the inner ear.

The main part of the inner ear is the cochlea, which converts mechanical oscillations into an electrical signal. In addition to the cochlea, the vestibular apparatus (see § 6.4) belongs to the inner ear, which is not related to the auditory function.

The human cochlea is a bone formation of circa 35 mm in length, it is a cone-shaped spiral with 2¾ turns. The diameter at the base is about 9 mm, the height is about 5 mm.

In fig. 8.8, the cochlea is sketchily shown expanded for a better view. Three canals pass along the cochlea. One of them begins from the oval window *8*, it is called *scala vestibuli*. The other canal comes from the round window *9*, it is called *scala tympani 10*. Scala vestibuli and scala tympani are connected in the area of the cochlear cupula by a small opening, helicotrema *11*. Thus, to some extent, both canals are an integral system filled with perilymph. Oscillations of the stapes *6* are transmitted to the oval window membrane *7*, and from it they go to perilymph, pushing out the round window membrane *9*. The space between scala vestibuli and scala tympani is called *cochlear canal 12*, and it is filled with endolymph. Between the cochlear canal and scala tympani, the basilar membrane *13* runs along the cochlea. On it, the organ of Corti is located containing receptor (hair) cells. The auditory nerve departs from the cochlea (in fig. 8.9 these details are not shown).

The organ of Corti (spiral organ) transforms mechanical oscillations into an electric signal.

The length of the basilar membrane is about 32 mm; it expands and gets thinner at the direction from the oval window at the top of the cochlea (from 0.1 to 0.5 mm). The basilar membrane is a very interesting structure for physics; it has frequency-selective properties. This drew the attention of Helmholtz, who considered the basilar membrane similar to a series of tuned strings of a piano The Nobel laureate Békésy that of this resonator theory was incorrect. Békésy showed that the membrane is a heterogeneous transmission line of mechanica excitation. Upon acoustic stimulation, a wave propagates along the basila membrane. Depending on the frequency, this wave attenuates differentl The lower the frequency, the farther away from the oval window the wave will propagate across the basilar membrane before it begins to attenuate. Thus before the beginning of attenuation, a wave with a frequency of 300 Hz will cover a distance of about 25 mm from the oval window, and a wave with a frequency of 100 Hz attains its maximum at a distance of about 30 mm.

On the basis of these observations, theories have been developed: perception of the pitch of a tone is determined by the position of maximum oscillations of the basilar membrane. Thus, in the inner ear, a certain functiona chain is traced: oscillations of the oval window membrane–perilymph oscillation–complex oscillations of the basilar membrane–stimulation of hair cells (organ of Corti receptors)–electric signal generation.

Some forms of deafness are associated with lesion of the receptor apparatus of the cochlea. In this case, the cochlea does not generate electric signals under the impact of mechanical oscillations. Such deaf people can be helped; to do this, it is necessary to implant electrodes into their cochlea and generat electrical signals corresponding to those that occur when a mechanical stimulus acts.

Such prosthetics of the main cochlea function (cochlear prosthetics) is developed in a number of countries. In Russia, cochlear prosthetics was developed and implemented at the Russian Medical University. The cochlea prosthetic appliances are shown in fig. 8.12, where *1* is the main frame, *2* is

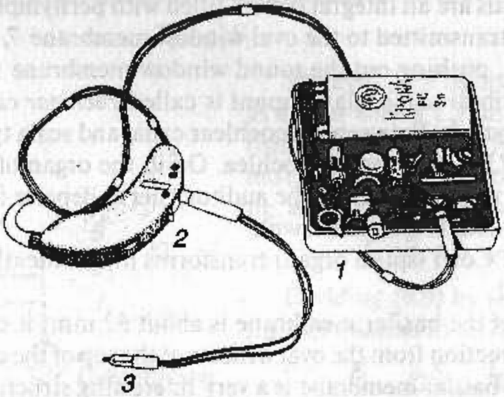


Fig. 8.12

the earpiece with microphone, *3* is the plug of the electrical connector for connection to the implantable electrodes.

§ 8.6. ULTRASOUND AND ITS APPLICATION IN MEDICINE

Ultrasound (US) is mechanical oscillations and waves with frequencies above 20 kHz.

The upper limit of ultrasonic frequencies can be conventionally considered 10^9–10^{10} Hz. This limit is determined by intermolecular distances and, therefore, depends on the aggregate state of the material in which the ultrasonic wave spreads.

Ultrasound radiators are devices used for generation of ultrasound. Electromechanical radiators based on the phenomenon of inverse piezoelectric effect are most widely used ones (see § 14.7). Inverse piezoelectric effect is mechanical deformation of bodies under the impact of electric field. The main part of such emitter (fig. 8.13, a) is a plate or rod *1* made from a material with well-pronounced piezoelectric properties (quartz, seignette salt, ceramic material based on barium titanate, etc.). Electrodes are placed on the plate surface in the form of conductive layers *2*. If alternating electrical voltage from generator *3* is applied to the electrodes, the plate, due to its reverse piezoelectric effect, begins to vibrate, emitting a mechanical wave of corresponding frequency.

The greatest effect of mechanical wave emission occurs when the resonance conditions are met (see § 7.6). Thus, for plates mm thick, the resonance occurs at the frequency of 2.87 MHz for quartz, at 1.5 MHz for seignette salt, and at 2.75 MHz for barium titanate.

An ultrasonic receiver can be developed on the basis of piezoelectric effect (direct piezoelectric effect). In this case, the action of a mechanical wave (ultrasonic waves) causes crystal deformation (fig. 8.13, b), which due to the piezoelectric effect generates an alternating electric field; the corresponding electrical voltage can be measured.

Application of ultrasound in medicine is based on the specifics of its propagation and characteristic properties. Let us consider this issue.

From the physical viewpoint, ultrasound is a mechanical (elastic) wave like any sound. Nevertheless,

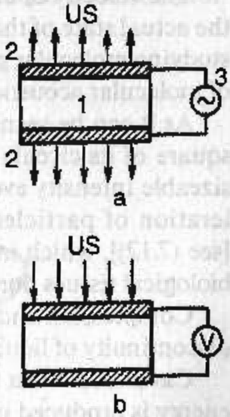


Fig. 8.13

ultrasonic wavelength is significantly less than the wavelength of the soun wave. Thus, in water wavelengths are 1.4 m (1 kHz, sound), 1.4 mm (1 MH ultrasound) and 1.4 μm (1 GHz, ultrasound). Wave diffraction (see § 24. depends on the ratio of the wavelength and the size of bodies on which the wave is diffracting, to a great extent. A "nontransparent" body with the size of 1 m will be no obstacle for the sound wave with a length of 1.4 m, but will becom an obstacle to an ultrasonic wave with a length of 1.4 mm, there will be an ultrasound shadow. In some cases, this permits neglecting the diffraction of ultrasonic waves in some cases, considering these waves as rays for refraction and reflection (similar to refraction and reflection of light rays).

Reflection of ultrasound at the interface of two media depends on the ratio of their wave resistance (see § 8.4). Thus, ultrasound is reflected well at the interface muscle–periosteum–bone, on the surface of hollow organs, etc. Therefor it is possible to determine the location and size of heterogeneous inclusions, cavities, viscera, etc. (ultrasonic location). In ultrasonic location, both continuou and pulsed radiation is used. In the first case, one studies a standing wave generated by interference of the incident wave, and the wave reflected from the interface. In the second case, one studies the echo pulse and measures the time of ultrasound traveling to the object under study, and back. Knowing the velocity of ultrasound propagation, the depth of the object location is determined.

Wave resistance of biological media is 3,000 times higher than the wave resistance of air. So, if the ultrasonic radiator is applied to the human body the ultrasound will not penetrate inside, but will be reflected due to the thin layer of air between the radiator and the biological object (see § 8.4). To eliminate the air layer, the surface of ultrasonic radiator is covered with a layer of oil

The velocity of ultrasonic wave propagation and its absorption depend on the actual state of the medium very much; this is the base of ultrasound use in studying molecular properties of matter. Studies of such a kind are the subjec of molecular acoustics.

As it can be seen from (7.53), intensity of the wave is proportional to the square of its circular frequency, so it is possible to obtain an ultrasound o sizeable intensity even with relatively small amplitude of oscillations. Acceleration of particles oscillating in the ultrasonic wave can also be high [see (7.12)], which indicates presence of sizeable forces acting on particles in biological tissues during ultrasonic irradiation.

Compression and vacuum created by ultrasound lead to formation produce discontinuity of liquid — cavitations.

Cavitations exist for a short time and collapse quickly; for all that, sizeable energy is produced in small spaces heating the material, as well as ionization and dissociation of molecules.

Physical processes conditioned by the effect of ultrasound cause the following main after-effects in biological objects:

- micro-vibrations at cellular and sub-cellular levels;
- destruction of biomacromolecules;
- restructuring and lesion of biological membranes, change in membrane permeability (see Chapter 13);
- heating effect.

Medical and biological applications of ultrasound can be divided into two basic trends: methods of diagnostics and research, and methods of treatment.

The first trend includes location methods and use of pulsed radiation. This is *echoencephalography* determining cerebral tumors and edema (fig. 8.14 shows Ekho-12 *echoencephalograph*); *ultrasonic cardiography* for dynamic measurement of heart dimensions; in ophthalmology it is *ultrasonic location* for determining the size of eye media. Ultrasonic Doppler effect is used to study the nature of motion of cardiac valves and measure the velocity of blood flow. For diagnostic purposes, the velocity of ultrasound determines the density of a knitted or damaged bone.

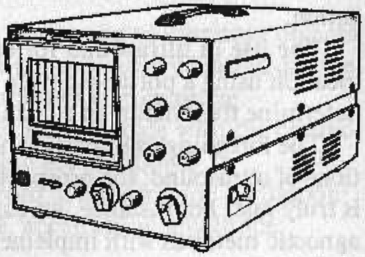


Fig. 8.14

The second trend is *ultrasound physiotherapy*. Fig. 8.15 shows a device used for this purpose, UTP-3M. The patient is irradiated by ultrasound using a special radiation head of the device. Usually, ultrasound with a frequency of 800 kHz is used for therapeutic purposes; its average intensity is about 1 W/cm^2 and less.

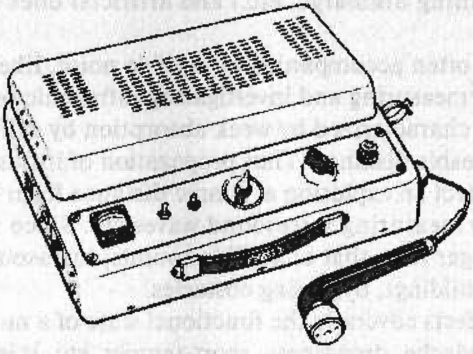


Fig. 8.15

The primary mechanism of ultrasonic therapy is mechanical and therma action on the tissue.

During surgery, ultrasound is used as an "ultrasonic scalpel" that can cut through both soft and osseous tissues.

The ability of ultrasound to break up bodies placed in liquid and creat emulsions is used in pharmaceutical industry for manufacture of medication Therapy of diseases such as tuberculosis, asthma, upper respiratory catarr makes use of aerosols of various medications produced with the help of ultrasound.

Nowadays, a new method of "welding" damaged or transplanted osseou tissues using ultrasound (*ultrasonic internal fixation*) is developed.

The destructive effect of ultrasound on microorganisms is used for sterilization.

The use of ultrasound for blind people is of interest. Due to ultrasoni location using a portable device "Orientir", it is possible to detect objects and determine their nature at a distance of 10 m.

The foregoing examples do not exhaust all medical and biological applications of ultrasound; the prospect of expanding the sphere of these applicatio is truly vast. For instance, we can expect emergence of fundamentally new diagnostic methods with implementation of ultrasonic holography in medicin (see Chapter 24).

§ 8.7. INFRASOUND

Infrasound is mechanical (elastic) waves with frequencies lower than thos perceived by a human ear (20 Hz).

Infrasound sources can be both natural objects and phenomena (sea, earthquake, lightning discharge, etc.) and artificial ones (explosions, cars, machines, etc.).

Infrasound is often accompanied by audible noise, like in a car, so there are difficulties in measuring and investigating infrasonic oscillations proper.

Infrasound is characterized by weak absorption by different media, so it spreads over a sizeable distance. Thus propagation of infrasound in the earth crust can help detect an explosion at a large distance from its source, foresee a tsunami etc. by measuring infrasound waves etc. Since the wavelength of infrasound is longer than that of audible sounds, infrasound waves diffract better and enter buildings, bypassing obstacles.

Infrasound affects adversely the functional state of a number of body systems: fatigue, headache, drowsiness, short-temper, etc. It is assumed that the primary mechanism of infrasound effect on the body is of resonant nature

Resonance occurs at close values of the driving force frequency and the frequency of free oscillations (see § 7.6). The frequency of free oscillations of the human body in a lying position is 3—4 Hz, in an upright position it is 5—12 Hz; the frequency of free oscillations of the thorax is 5—8 Hz, in the abdomen it is 3—4 Hz, etc.; thus, the frequencies correspond to infrasound ones.

Reducing infrasound intensity level in residential, industrial and transport premises is a task of hygiene.

§ 8.8. VIBRATIONS

In technology, mechanical oscillations of various structures and machines are called *vibrations.*

They have an impact on the human who is exposed to vibrating objects. This impact may be harmful and in certain conditions cause vibration disease, or have a useful, therapeutic effect (vibrotherapy and vibromassage).

The main physical characteristics of vibrations coincide with the characteristics of mechanical oscillations of bodies and are as follows:

- oscillation frequency or harmonic spectrum of anharmonic oscillation;
- oscillation amplitude, velocity amplitude, acceleration amplitude;
- energy and average power of oscillations.

In addition, to understand the effects of vibrations on a biological object, it is important to understand the propagation and attenuation of oscillations in a body. For studying of this problems, models consisting of inertial masses, elastic and viscous elements are used (see § 10.3).

Vibrations are the source of audible sound, ultrasound, and infrasound.

Chapter 9

Flow and properties of liquids

Liquids are materials that due to their properties occupy the intermedia position between gases and solids. Liquid media constitute the most par of the body, their motion ensures metabolism and supply of oxygen to cells, so mechanical properties and flow of liquids are of particular interes to physicians and biologists.

The material covered in the chapter is related to hydrodynamics which is a branch of physics, which studies the motion of incompressible liquids and their interaction with surrounding solids, and to rheology, the study of deformation and fluidity of matter.

§ 9.1. VISCOSITY OF LIQUID. NEWTON EQUATION. NEWTONIAN AND NON-NEWTONIAN LIQUIDS

During the flow of real liquid, its individual layers affect one another wi forces tangential to the layers. This phenomenon is called ***internal frictio*** or ***viscosity.***

Let us consider flow of a viscous liquid between two solid plates (fig. 9.1), of which the lower one is fixed, and the upper one moves with velocity υ_U. Conventionally, we present the liquid in the form of several layers 1, 2, 3 etc. The layer adhering to the bottom is steady too. Moving away from the bottom (the lowe plate) the liquid layers have a higher velocit ($\upsilon_1 < \upsilon_2 < \upsilon_3 < \ldots$ etc.), the maximum veloci ty υ_U will be in the layer adhering to the upper plate.

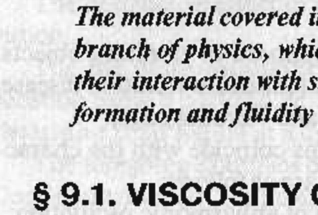


Fig. 9.1

Layers interact one with another. Thus, the third layer tends to accelerate the motio of the second one, but it is inhibited by the second one, while is accelerated by the fourth layer, etc. The internal frictio force is proportional to area S of the interacting layers and is the higher, the higher their relative velocity is.

Since separation into layers is conventional, the force is conventionally e pressed depending on change in velocity related to the length at directio which is perpendicular to the velocity, i.e., $\mathrm{d}\upsilon/\mathrm{d}x$ is the ***velocity gradient*** (***shear rate***):

$$F_{\mathrm{fr}} = \eta \cdot \frac{\mathrm{d}\upsilon}{\mathrm{d}x} \cdot S. \tag{9.1}$$

This is the ***Newtonian equation.*** In this, η is coefficient of proportionality that is called coefficient of internal friction or ***dynamic viscosity*** (or simply *viscosity*). Viscosity depends on the state and molecular properties of liquid (or gas).

The unit of viscosity is *Pascal-second* (Pa · s). In the CGS system, viscosity is expressed in *poises* (P): 1 Pa · s = 10 P.

For many liquids, viscosity is independent from the velocity gradient; such liquids meet the Newtonian equation (9.1) and are called *Newtonian* ones. Liquids that do not meet the equation (9.1) are *non-Newtonian* ones. Sometimes the viscosity of Newtonian liquids is called ***normal***, and that of non-Newtonian liquids is called ***irregular***.

Liquids consisting of complex and large molecules, such as polymer solutions, and spatial structures formed due to adhesion of molecules or particles, are non-Newtonian. Their viscosity, other things being equal, is much higher than that of simple liquids.

Viscosity can increase because when these liquids flow, the work of external forces is expended not only on overcoming true Newtonian viscosity, but also onto destroying their structure. Blood is a non-Newtonian liquid.

§ 9.2. VISCOUS LIQUID FLOW THROUGH PIPES. POISEUILLE FORMULA

The viscous liquid flow through the pipes is of particular interest for medicine since the circulatory system consists mainly of cylindrical vessels of different diameters.

Due to symmetry, it is clear that in a pipe, the particles of the flowing liquid, equidistant from the axis, have the same velocity. The particles moving along the pipe axis have the highest velocity; the liquid layer closest to the pipe wall is immobile.

Approximate distribution of velocity of liquid particles in the pipe cross-section is shown in fig. 9.2.

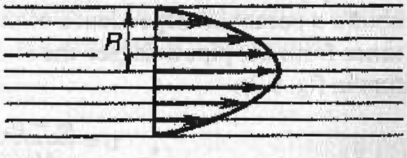


Fig. 9.2

To determine the dependency $\upsilon = f(r)$ let us select a cylindrical volume of liquid of radius r and length l (fig. 9.3, a).

At the bases of this cylinder pressures p_1 and p_2 are maintained respectively which produces the resultant force

$$F = p_1\pi r^2 - p_2\pi r^2 = (p_1 - p_2)\pi r^2. \tag{9.2}$$

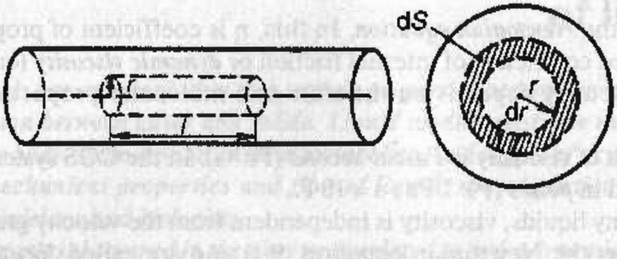


Fig. 9.3

An internal friction force acts on the lateral surface of the cylinder from the surrounding liquid layer, and it is equal to [see (9.1)]

$$F_{fr} = \eta\frac{d\upsilon}{dx}S = \eta\frac{d\upsilon}{dr}2\pi rl. \tag{9.3}$$

where $S = 2\pi rl$ is the area of the lateral surface of the cylinder. Since the liquid flow is uniform then forces acting onto the cylinder selected are equal: $F = F_{fr}$. Substituting (9.2) and (9.3) into this equation we obtain

$$(p_1 - p_2)\pi r^2 = -\eta\frac{d\upsilon}{dr}2\pi rl. \tag{9.4}$$

The "–" sign in the right-hand side of the equation is due to the fact that $d\upsilon/dr < 0$ (velocity decreases with increasing r). From (9.4) we have

$$d\upsilon = -\frac{p_1 - p_2}{2l\eta}r\,dr.$$

Let us integrate this equation:

$$\int_0^{\upsilon} d\upsilon = -\frac{p_1 - p_2}{2l\eta}\int_R^r r\,dr, \tag{9.5}$$

in this, the lower limits correspond to the layer adhering to the inner surface of the pipe ($\upsilon = 0$ for $r = R$); the upper limits are variables. Solving (9.5), we obtain a parabolic dependence of the velocity of the liquid layers on their distance from the pipe axis (see the curve that envelopes the ends of velocity vectors in fig. 9.2):

$$\upsilon = \frac{p_1 - p_2}{4l\eta}(R^2 - r^2). \tag{9.6}$$

The layer flowing along the pipe axis has the highest velocity ($r = 0$):

$$\upsilon_{\max} = (p_1 - p_2)R^2/(4l\eta).$$

Let us find out on what factors volume Q of liquid flowing through a horizontal pipe over 1 s depends. For this, let us select a cylindrical layer of radius r and thickness dr. The cross-section of this layer is (fig. 9.3, b) $\mathrm{d}S = 2\pi r\,\mathrm{d}r$. Since the layer is thin, it can be assumed that it moves with the same velocity υ. Over 1 s, the layer carries the volume of liquid

$$\mathrm{d}Q = \upsilon\,\mathrm{d}S = \upsilon \cdot 2\pi r\,\mathrm{d}r. \tag{9.7}$$

Substituting (9.6) into (9.7), we obtain

$$\mathrm{d}Q = \pi\frac{p_1 - p_2}{2l\eta}(R^2 - r^2)r\,\mathrm{d}r,$$

whence integrating across the entire cross-section we obtain

$$Q = \pi\frac{p_1 - p_2}{2l\eta}\int_0^R (R^2 - r^2)r\,\mathrm{d}r = \frac{\pi R^4}{8\eta}\frac{p_1 - p_2}{l}. \tag{9.8}$$

This dependence is known as ***Poiseuill's formula***.

As it can be seen from (9.8), under specified external conditions (p_1 and p_2), the more amount of liquid flows through the pipe, the lower its viscosity and the greater the radius of the pipe. The strong dependence of Q on the radius is due to a change not only in volume, but also in the relative proportion of layers located near the pipe surface.

Let us draw an analogy between Poiseuill's formula (9.8) and Ohm's law for a circuit section without a current source. The difference of potentials corresponds to the pressure difference at the ends of the pipe, the strength of current corresponds to the volume of liquid flowing through the pipe section in 1 second, electrical resistance corresponds to ***hydraulic resistance***:

$$X = 8\eta l/\pi r^4. \tag{9.9}$$

The greater viscosity η, the greater the hydraulic resistance and length l of the pipe, and the smaller the cross-section area. In some cases, the analogy between electric and hydraulic resistance makes permits using the rule of finding electric resistance for series and parallel connection of conductors for determining the hydraulic resistance of a system of serial or parallel connection of pipes. Thus, the total hydraulic resistance of three pipes connected in series (fig. 9.4, a) or in parallel (fig. 9.4, b), is calculated by the formulas:

$$X = X_1 + X_2 + X_3, \tag{9.10}$$

$$X = \left(\frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3}\right)^{-1}. \tag{9.11}$$

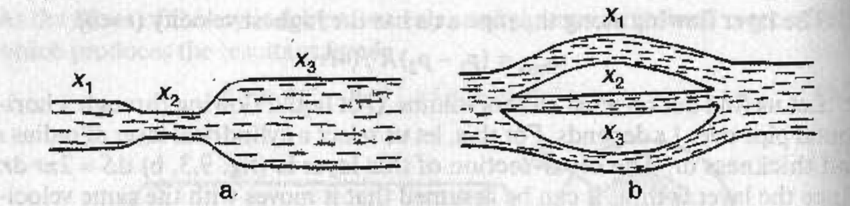


Fig. 9.4

To generalize the Poiseuille equation for pipes of variable cross-section, let us replace $(p_1 - p_2)/l$ with pressure gradient $\mathrm{d}p/\mathrm{d}l$ and we obtain

$$Q = \frac{\pi R^4}{8\eta} \cdot \frac{\mathrm{d}p}{\mathrm{d}l}. \tag{9.12}$$

Let us install manometric tubes in different places of a horizontal cylindr cal pipe of varying cross-section with flowing viscous liquid (fig. 9.5, a). The show that static pressure along the pipe of varying cross-section decreas proportionally to l; $\mathrm{d}p/\mathrm{d}l = \mathrm{const}$. Since Q is the same [see (9.12)], the pressu gradient is greater in pipes of smaller radius. The plot curve of pressure dependence on the distance along the pipes is approximately shown in fig. 9.5, b.

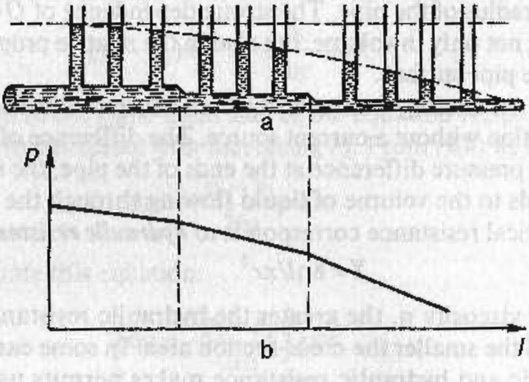


Fig. 9.5

§ 9.3. MOTION OF BODIES IN VISCOUS LIQUID. STOKES LAW

Viscosity is manifested not only when liquid passes through vessels, but also when bodies move in a liquid. When velocities are low, in accordan with Newton's equation, the resistance force on a moving body is proportion to viscosity of a liquid, velocity of the body and depends on the size of the body. Since it is impossible to specify a general formula for a resistance force, we will limit ourselves to considering a particular case.

The simplest shape is a sphere. For a spherical body (a ball), the Stokes law says that resistance force during its motion in a vessel with liquid depends on the foregoing factors:

$$F_{\text{fr}} = 6\pi\eta r\upsilon, \qquad (9.13)$$

where r is the ball radius; υ is velocity of movement. This law is obtained under the assumption that vascular walls do not affect the motion of the body.

When the ball falls in a viscous medium (fig. 9.6), it is subject to three forces:

a) gravity force $mg = 4/3\rho\pi r^3 g$ (ρ is the ball's density);

b) buoyancy (Archimedean) force $F_A = m_1 g = 4/3\rho_1\pi r^3 g$, where m_1 is the mass of liquid displaced by the ball; ρ_1 is the liquid density;

c) F_{fr} is resistance force calculated by formula (9.13).

When the ball gets to a viscous liquid, its velocity decreases. Since resistance force is directly proportional to velocity, the velocity will be decreasing until motion becomes uniform. In this case (see fig. 9.6),

$$m\mathbf{g} + \mathbf{F}_A + \mathbf{F}_{\text{fr}} = 0,$$

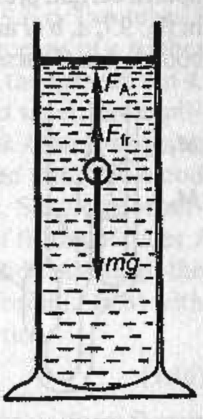


Fig. 9.6

or in scalar form, after substituting corresponding expressions for the forces

$$\tfrac{4}{3}\pi r^3\rho g - \tfrac{4}{3}\pi r^3\rho_1 g - 6\pi\eta r\upsilon_0 = 0, \qquad (9.14)$$

where υ_0 is the velocity of uniform motion (falling) of the ball. From (9.14), we obtain:

$$\upsilon_0 = 2(\rho - \rho_1)r^2 g/(9\eta). \qquad (9.15)$$

Formula (9.15) is valid for the ball motion not only in a liquid, but also in a gas. It can be applied, in particular, to calculating the time of dust deposition in air. Let us explain this with the following example. The viscosity of air, which is a medium containing various dust particles, is $\eta = 0.000175$ P·s. About 80% of dust, discovered in the lungs of dead people, are particles from 5 to 0.2 μm. If the dust particles are assumed to be spherical and the dust density to be equal to the density of soil ($\rho = 2.5$ g/cm^3), then, calculating the velocity of falling of these dust particles by formula (9.15), we find that the velocities are in the range of 0.2–0.0003 cm/s. For complete deposition of such dust in a room 3 m high, it will take about 12 days, provided the air is absolutely immobile and there is no Brownian motion.

§ 9.4. METHODS DETERMINING LIQUID VISCOSITY. CLINICAL METHOD OF BLOOD VISCOSITY ASSESSMENT

The totality of methods for measuring viscosity is called *viscosimetry*, and instruments used for these purposes are called *viscometers*. Let us consider the most common methods of viscosimetry.

The capillary method is based on the Poiseuille formula and consists in measuring the time of flow of a liquid with known mass under the impact of gravity upon a certain pressure drop. Capillary viscometers of different types are shown in fig. 9.7, a, b (*1* are measurement reservoirs; M_1 and M_2 are marks showing the boundaries of these vessels; *2* are capillaries; *3* are receiving vessels).

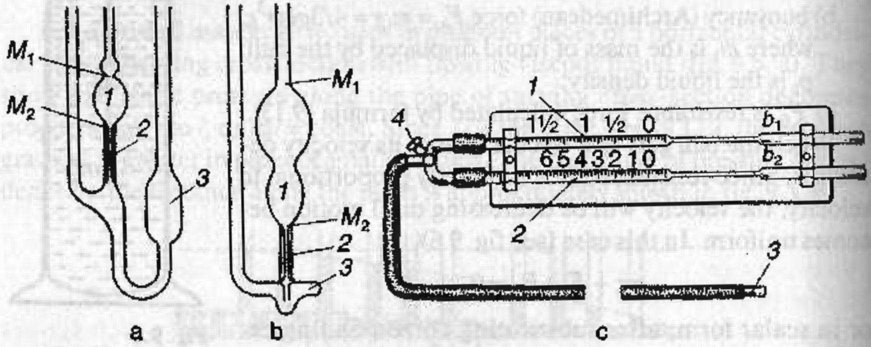


Fig. 9.7

A capillary viscometer is used to assess blood viscosity.

Capillary viscometers make it possible to measure viscosity from 10^{-5} Pa · s (gases) to values of 10^4 Pa · s (grease lubricant).

The falling ball method is used in viscometers based on Stokes law. From formula (9.15) we find

$$\eta = 2(\rho - \rho_l)r^2 g/(9\upsilon_0).$$

Thus, knowing the values of the right part of this formula and measuring the velocity of uniform fall of the ball, you can find the viscosity of the liquid.

The limit of effective range of viscometers with a moving ball is $6 \cdot 10^4$–250 Pa · s.

Rotary viscometers are also used; in them, the liquid is in a gap between two coaxial bodies, for instance, cylinders. One cylinder (rotor) rotates, and the other one is fixed. Viscosity is measured by angular velocity of the rotor, which creates a certain moment of force on the fixed cylinder, or by the moment of force acting on the stationary cylinder for a given angular rotation velocity of the rotor.

Rotary viscometers make it possible to determine viscosity of liquids in the range of 1–10^5 Pa·s, that is lubricating oils, molten silicates and metals, high viscosity varnishes, glues, mud, etc.

Rotary viscometers permit changing the velocity gradient by setting different rotor angular rotation velocities. This allows measuring viscosity for different velocity gradients and detrermining the dependence $\eta = f(d\upsilon/dx)$, which is typical of non-Newtonian liquids.

Nowadays in clinical settings, to determine blood viscosity, Hess ***viscometer*** with two capillaries is used. Its diagram is shown in Fig. 9.7, c. Two tubes *1* and *2* connect two identical capillaries a_1b_1 and a_2b_2. By means of a rubber bulb or taking air by mouth through tip *3*, by turns, through the tee with tap *4*, capillary a_1b_1 and tube *1* are filled up to mark *0* with distilled water, and capillary a_2b_2 and tube *2* are filled with blood sample until mark *0*. After that, using the same methods simultaneously, both liquids are shifted until the blood reaches mark *1*, and water reaches another mark in its tube. Since the conditions for water and blood flow are identical, the volumes of filling of tubes *1* and *2* will be different to different to viscosity of these liquids. Although the blood is a non-Newtonian liquid, we use the Poiseuille formula (9.8) with a certain approximation and write down the obvious proportion:

$$Q_w / Q_b = \eta_b / \eta_w. \tag{9.16}$$

Taking into account that the total volume *V* of a liquid in uniform flow is connected to *Q* by formula $V = Qt$, where *t* is time, instead of (9.16), we obtain

$$V_w : V_b = \eta_b : \eta_w,$$

where V_b is the volume of blood in tube *2* from mark *0* to mark *1*; V_w is the volume of water in tube *1* from mark *0* to mark obtained by measurement; η_b and η_w are the viscosity of blood and water, respectively. At the same temperature, the ratio of blood to water viscosity is called ***relative blood viscosity***.

In a Hess viscometer, the volume of blood is constant, and the volume of water is determined by marks on tube *1*, so relative viscosity of the blood is obtained directly. For convenience of quantification, cross-sections of tubes *1* and *2* are made to be different, so, despite different volumes of blood and water, their levels in the tubes will be approximately equal.

Normal human blood viscosity is 4–5 mPa·s; in pathological cases, it ranges from 1.7 to 22.9 mPa·s, which affects erythrocyte sedimentation rate (ESR). Venous blood has a slightly higher viscosity than arterial blood. In hard physical work, blood viscosity increases. Some infectious diseases increase blood viscosity, while other diseases, such as typhoid and tuberculosis, reduce it.

§ 9.5. LAMINAR AND TURBULENT FLOW. REYNOLDS NUMBER

The liquid flow discussed above is stratified or laminar. An increase in the velocity of a viscous liquid flow, due to the heterogeneous pressure along th cross-section of the pipe, generates a vortex area; and the motion becom ***vortical***, or ***turbulent***. For turbulent flow, the velocity of particles in each plac changes continuously and chaotically, the motion is unsteady.

The nature of liquid flow through a pipe depends on the liquid propert velocity of its flow, pipe sizes, and is determined by the *Reynolds number*:

$$\mathrm{Re} = \rho_l \upsilon D/\eta,$$

where ρ_l is liquid density; η is its viscosity; D is the pipe diameter; υ is the flow velocity.

If Reynolds number is greater than a certain critical value ($\mathrm{Re} > \mathrm{Re}_{cr}$) th the liquid flow is turbulent. For instance, for smooth cylindrical tube $\mathrm{Re}_{cr} \approx 2{,}300$.

Since the Reynolds number depends on the liquid viscosity and densit it is convenient to introduce their ratio, called ***kinematic viscosity***:

$$\nu = \eta/\rho_l.$$

Using this concept, the Reynolds number can be expressed as

$$\mathrm{Re} = \upsilon D/\nu. \tag{9.17}$$

The unit of kinematic viscosity is *square meter per second* (m^2/s), in the CGS system, it is *stokes* (St); their ratio is $1\ \mathrm{St} = 10^{-4}\ m^2/s$.

Kinematic viscosity takes into account the effect of internal friction on the nature of the liquid flow to a greater extent than dynamic viscosity. Thus the viscosity of water is about 100 times higher than that of air (at 0 °C), bu the kinematic viscosity of water is 10 times less than that of air, so viscosity affects the nature of air flow stronger than water flow.

As can be seen from (9.17), the nature of liquid flow depends on pipe dimensions very much. In wide pipes, even at relatively low velocities, turbulen flow can occur. Thus, in a tube with a diameter of 2 mm, the water flow becomes turbulent at a velocity more than 127 cm/s, and in a pipe with a diameter of 2 cm, it occurs at a velocity of about 12 cm/s (at a temperature 16 °C). The blood flow through such a tube would be turbulent at a velocity of 50 cm/ but in fact, in blood vessels with a diameter of 2 cm turbulent flow occurs ev at a lower velocity.

In arteries, the blood flow is normally laminar; slight turbulence occur near the valves.

If there is an abnormality, when the viscosity is below normal, the Reynolds number can exceed its critical value; the flow becomes turbulent.

Turbulent flow generates additional energy expenditure because of the excessive liquid flow, which in our case (blood) leads to additional heart work. The noise occurring in turbulent blood flow can be an indicator to diagnose. This noise is heard on the brachial artery when taking blood pressure.

Airflow in the nasal cavity is normally laminar. However, in inflammation or any other disorder, it can become turbulent, which will entail additional work of the respiratory muscles.

Reynolds number is a similarity criterion. For simulation of hydro- and aerodynamic systems, blood circulation in particular, the model shall have the same Reynolds number as the real system; otherwise, there will be no correspondence between them. This also applies to simulating flow around a body moving in a fluid.

From (9.17) it is clear that reduction of a model size compared to the real system can be balanced by increasing flow velocity or decreasing kinematic viscosity of model fluids.

§ 9.6. SPECIFICS OF MOLECULAR STRUCTURE OF LIQUIDS

Common liquids are isotropic; they are amorphous solids structurally. Liquids have show an ordered relative arrangement of the nearest particles. The distances between the molecules are small; the forces of interaction are sizeable, which generates low compressibility of liquids: a small decrease in the distance between molecules causes large forces of intermolecular repulsion. Like solids, liquids are hardly compressible and have a high density; like gases, liquids take the form of a vessel in which they are. These properties of liquids are due to specifics of their molecules thermal motion. In gases, molecules move chaotically, progressively on small segments of their path, there is no order in particle arrangement. In crystal bodies, the particles oscillate around certain equilibrium positions: points of their crystal lattice. According to Frenkel's theory, liquid molecules, like solid particles, oscillate around equilibrium positions, but the latter are not constant. In a certain time, called settled life, the molecule jumps to a new equilibrium position at a distance equal to the average distance between neighboring molecules.

Let us calculate average distance δ between molecules in a liquid. Since $\delta_3 \approx 1/n$, where $n = N_A \cdot \rho / M$ is molecule concentration in the liquid (N_A is the Avogadro constant; ρ is liquid density; M is the molar mass), then

$$\delta = 1/\sqrt[3]{n} = \sqrt[3]{M/(N_A \cdot \rho)}. \tag{9.18}$$

The order of δ is 10^{-10}; e.g., for water $\delta \approx 3 \cdot 10^{-10}$ m.

The average time of settled life of the molecule is called the time of relaxation r. When temperature increases and pressure drops, relaxation time reduces to a great extent, which causes higher mobility of the liquid molecules and its low viscosity.

In order for a liquid molecule to jump from one equilibrium state to another, the bonds with the next molecules should be broken forming bonds with new neighbors. The process of breaking bonds requires energy expenditure E_a (*activation energy*), released during formation of new bonds. Such molecule transition from one position of equilibrium to another is ***transition over a potential barrier*** with a height of E_h. The molecule receives energy to overcome the potential barrier at the expense of thermal motion energy of neighboring molecules. The dependence of relaxation time on the liquid temperature and activation energy is expressed by the formula following from the Boltzmann distribution:

$$r = r_0 e^{E_a/(kT)}, \tag{9.19}$$

where r_0 is the average molecule oscillation period about the equilibrium position.

Knowing average displacement δ and average time r, it is possible to determine average velocity of molecules in a liquid:

$$\upsilon = \delta/r = (\delta/r_0)e^{E_a/(kT)}. \tag{9.20}$$

This velocity is low in comparison to average velocity of molecules motion in gases. Thus, at the same temperature, velocity of water molecules is 20 times lower than that of vapor.

§ 9.7. SURFACE TENSION

On interfaces of a liquid and its saturated vapor, of two immiscible liquids, a liquid and solid, a force occurs conditioned by different intermolecular interaction of the neighboring media.

Each molecule located inside a space of liquid is uniformly surrounded by neighboring molecules and interacts with them, but the resultant of these forces is zero. Due to the heterogeneity of the environment, there is a force not compensated by action of other molecules in the liquid. The force acts on a molecule located near the interface of two media. Therefore, to move molecules from inside to the surface layer, it is necessary to produce work.

Surface tension is determined by the ratio of work expended on creation of a liquid surface at a constant temperature, to the area of the surface:

$$\sigma = A/S. \tag{9.21}$$

The condition of stable equilibrium of liquids is the minimum energy of the surface layer, so in absence of external forces or in a state of weightlessness, the liquid tends to have a minimum surface area at the given volume and takes a spherical shape.

Surface tension can be determined not only energetically. The tendency of the upper layer of a liquid to reduce is due to the presence of tangential forces in this layer — the forces of surface tension. If you select on the liquid surface a certain segment l long (fig. 9.8), it is possible to represent these forces conventionally with arrows which are perpendicular to the segment.

Surface tension is equal to the ratio of the surface tension force to the length of the segment on which this force is applied:

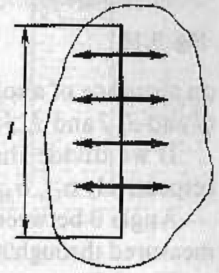


Fig. 9.8

$$\sigma = F/l. \tag{9.22}$$

It is known from the school physics course that definitions (9.21) and (9.22) are identical. We present the values of surface tension for some liquids at 20 °C (Table 9.1).

Table 9.1

Liquid	σ, N/m	Liquid	σ, N/m
Water	0.0725	Mercury	0.47
Bile	0.048	Alcohol	0.022
Milk	0.05	Blood serum	0.06
Urine	0.066	Ether	0.017

Surface tension depends on temperature. Far from the critical temperature, its value decreases linearly upon an increase in temperature. Reduction of surface tension can be attained by additing surfactants to the liquid, which reduces the energy of the surface layer.

§ 9.8. WETTING AND NON-WETTING. CAPILLARY PHENOMENA

On the interface of different media, ***wetting*** or ***non-wetting*** is present.

Let us consider the behavior of a liquid drop falling onto the surface of another liquid without mixing with it (fig. 9.9) and a drop of liquid

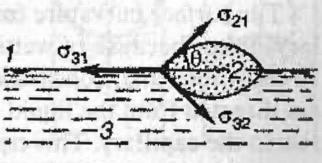


Fig. 9.9

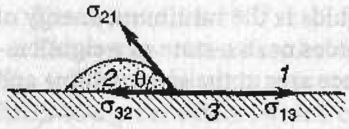


Fig. 9.10

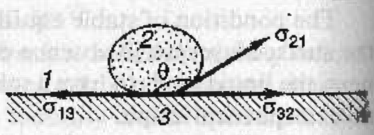


Fig. 9.11

on a surface of a solid (fig. 9.10 and 9.11). On the interfaces of each two medi (*1* and *3*, *2* and *1*, *3* and *2*) surface tension forces are active.

If we divide these forces by the circumference of the drop, we obtai respectively σ_{13}, σ_{21}, σ_{32}.

Angle θ between the surface wetted and tangent to the surface of the liqui measured through it is called *boundary angle*.

The following value is taken as a measure of wetting:

$$\cos\theta = (\sigma_{32} - \sigma_{13}/\sigma_{21}). \qquad \text{(9.2}$$

If $\sigma_{32} > \sigma_{13}$ (fig. 9.10), that is, the forces of interaction between molecules o liquid and a solid are greater than the forces of interaction between the soli and gas molecules, then $\theta < \pi/2$ and the liquid wets the solid; in this case, its surface is referred to as ***hydrophilic***. When of $\sigma_{32} < \sigma_{13}$ (fig. 9.11) $\theta > \pi/2$, the liquid does not wet the body; in that case, its surface is called ***hydrophobi*** Non-wetting liquid does not flow through small holes in a solid. In case o $\sigma_{32} - \sigma_{13} = \sigma_{21}$, intermolecular forces are fully equilibrated ($\theta = 0$). In thi case, no equilibrium is possible and the drop spreads over the surface of a solid until it covers the entire surface or a monomolecular layer is forme This case is ***ideal wetting.*** Spreading of alcohol or water on a clean glass su face, oil spill on water, etc. can be cited as an example with a certain approxi mation.

Under action of surface tension forces, the surface layer of the liqui becomes curved and exerts additional pressure with respect to the externa pressure of Δp. The surface layer is like an elastic shell, for example, a rubbe film. The resultant of surface tension forces of the curved surface is directe towards the concave side (to the center of curvature). In case of a spheric surface whose radius of curvature is r, additional pressure is:

$$\Delta p = 2\sigma/r. \qquad (9.24)$$

The surface curvature (meniscus), occurs, in particular, in narrow (capillary) tubes because of wetting or non-wetting of their surface by a liquid. In case of wetting, concave meniscus is formed (fig. 9.12). The pressure forces are directed from the liquid outward, that is, upward, and cause the liquid to rise in the capillary. This equilibrium state shown in the figure, occurs whe pressure $\rho g h$ counterbalances Δp.

From fig. 9.12, it is clear that $r = R/\cos\theta$, where R is the capillary radius. Thus, [see (9.24)] we obtain

$$\Delta p = 2\sigma\cos\theta/R. \qquad (9.25)$$

Then

$$\rho g h = 2\sigma\cos\theta/R,$$

whence, the height of the capillary is

$$h = 2\sigma\cos\theta/(R\rho g), \qquad (9.26)$$

that is, it depends on the properties of a liquid and capillary material, as well as its radius.

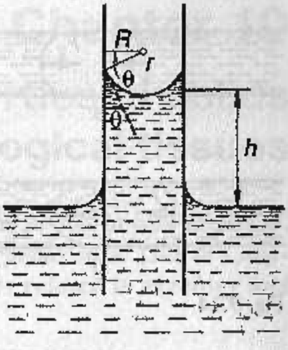


Fig. 9.12

In case of non-wetting, $\cos\theta < 0$, and formula (9.26) will yield the height of liquid dropping in the capillary.

Capillary phenomena determine conditions of vapor condensation, liquid boiling, crystallization, etc. Thus a vapor molecule (fig. 9.13, point A) over a concave meniscus of liquid is influenced by more liquid molecules, and, therefore, a greater force than in case of a convex meniscus. This follows quite clearly from fig. 9.13, in which the dotted line shows the conventional sphere of molecular action, and the dash line shows volumes of the liquid whose molecules are attracting the selected molecule of vapor. This results in capillary condensation in thin tubes wetted even at relatively low air humidity. Thanks to this, porous materials can hold a sizeable amount of liquid from vapor, which leads to wetting of linen, cotton wool in damp rooms, makes drying of hygroscopic solids difficult, helps in retaining moisture in the soil, etc. On the contrary, non-wetting liquids do not penetrate porous bodies. This fact explains water impenetrability of bird feathers covered in grease.

Let us consider the behavior of an air bubble in a capillary filled with liquid. If the pressure of the liquid onto the bubble is equal on all sides, both menisci of the bubble will have the same radius of curvature (fig. 9.14, a). At excessive pressure from one side, e.g., when the liquid is flowing, the menisci deform, their radii of curvature change (fig. 9.14, b), additional pressure Δp becomes unequal on different sides. All this will lead to such an impact onto the liquid from the air (gas) bubble which will hinder or stop the liquid motion. **Such phenomena can occur in a human circulatory system.**

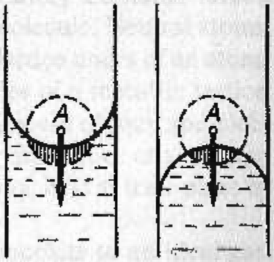


Fig. 9.13

Air bubbles, which have got in to blood, can block a small vessel and deprive the organ of blood supply. This phenomenon, called *embolism*, can lead to a serious functional disorder or even death. For instance, air embolism can occur when large veins are injured: the air that has come into the blood stream forms an air bubble that obstructs passage of blood. Air bubbles should not get in to a vein at intravenous infusions.

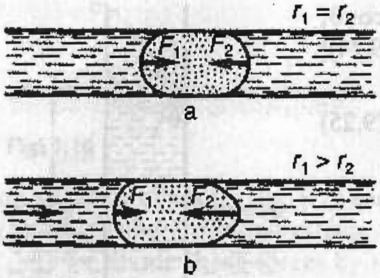


Fig. 9.14

Gas bubbles in blood can appear in divers when rising quickly from a great depth to the surface, pilots and astronauts upon a loss of cabin pressure or spacesuit at high altitude (gas embolism). This is due to the fact that blood gases pass from dissolved state to free gaseous one because of a lower ambient atmospheric pressure. Upon a decrease in pressure, formation of gas bubbles is mostly due to nitrogen since it accounts for the greatest portion of the total pressure of gases in the blood and does not participate in gas exchange between the body and atmospheric air.

Chapter 10

Mechanical properties of solids and biological tissues

A characteristic indicator of a solid body is the ability to hold its shape. Solids can be divided into crystalline and amorphous ones. Like in Chapter 9, the information under consideration is related to rheology and biorheology.

§ 10.1. CRYSTALLINE AND AMORPHOUS BODIES. POLYMERS

A distinctive indicator of the crystal state is ***anisotropy***, the property of being directionally dependent, which implies different properties (mechanical, thermal, electrical, optical ones) in diffferent directions.

Anisotropy of crystals is due to an ordered arrangement of atoms or molecules from which they are built, manifested in the correct external faceting of individual single crystals. However, as a rule, crystalline bodies are encountered in the form of polycrystals which are aggregates of individual small crystals (crystallites) fused together and randomly oriented. In this case, anisotropy is present within the crystallites.

The arrangement of atoms (molecules) of a crystal is ordered because they are located in the nodes of geometrically regular structures, forming a *crystal* (spatial) *lattice*. Depending on the nature of particles in the nodes and the nature of interaction forces, there are four types of crystal lattices: ***ionic***, ***atomic***, ***metallic***, ***and molecular***.

In nodes of the crystal lattice of an ionic crystal, there are ions of different signs. The forces of interaction between them are mainly Coulomb forces. Such a crystal is integrally considered to be a single molecule. Neutral atoms, between which covalent bonds are active, occupy the lattice nodes of an atomic crystal. Positive metal ions are located in all nodes of a metallic lattice. Among them, the electrons move chaotically. The system of ions and electrons produces a metallic bond. In the nodes of the crystal lattice of a molecular crystal, there are molecules oriented in a certain way, held at their place by forces of intermolecular interaction.

From the energy viewpoint, an ideal crystal is opposite to an ideal gas. In the latter, the absolute value of interaction energy is much less than kT, which is the average energy of chaotic thermal motion. On the contrar in a crystal, due to powerful interaction forces, the absolute value of intera tion energy is much higher than kT. Therefore, thermal motion in the crysta cannot destroy the bonds between particles; hence, they make small oscill tions near equilibrium positions. In a crystal, interaction among any particl is expressed by dependence of potential energy E_p on distance r between the (fig. 10.1). The plot curve is not symmetric with respect to the minimu straight line. Distance r_0 between interacting particles corresponds to t minimum value of potential energy at $T = 0$ K. Let at temperature T_1 the to (kinetic and potential) energy be E_1. This means that the particle oscillat between two points A_1 and B_1. The average distance between two particle $r_1 = (|0A_1| + |0B_1|)/2$. At $T_2 > T_1$ the particle energy is equal to $E_2 > E_1$ and it oscillates between A_2 and B_2. The average distance between the particles is $r_2 = (|0A_2| + |0B_2|)/2$. Since the potential energy curve is asymmetri the average distance between the particles increases in proportion to heatin $r_0 < r_1 < r_2 < r_3 < \ldots$ at $0\text{ K} < T_1 < T_2 < T_3 < \ldots$, which exactly causes therma expansion of bodies.

The main macroscopic feature of amorphous solids consists in natura isotropy of their properties and absence of a definite melting point, whic is due to their internal structure.

The distinctive feature of the internal structure of solids in amorphou state is absence of long-range order, which is typical of crystal state, that is, strict reiteration in the arrangement of atoms or groups of atoms in all direc tions along the entire body.

At the same time, a material in its amorphous state has a short-range orde that is, some order in arrangement of adjacent particles. In proportion of distance, this order decreases.

Having a lower order of their internal structure, under the same conditions, amorphous solids have a greater specific volume, entropy, and intern energy than crystals.

These bodies have a sufficient equi librium state only at high temperatu and low pressure, which is associate with achieving a certain arrangement particles and distances between the In accordance with this, depending o the rate of external influence, amor phous solids can be elastic or fluid. For instance, if a piece of black pitch is in a vessel, in a long time interval, it wil

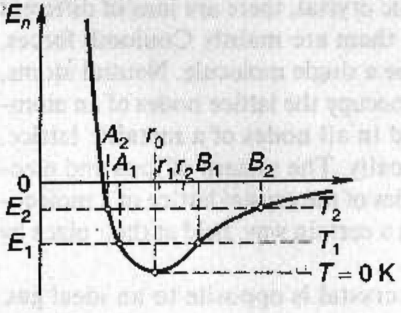


Fig. 10.1

take the shape of the vessel, that is, the pitch will show properties of fluidity. If you hit this piece with a hammer, it will split like a fragile solid.

Amorphous state is native to materials of various chemical nature. At low pressure and high temperature, materials in this state are very mobile: low-molecular ones are liquids; high molecular components are in highly elastic state. At decreasing temperature and increasing pressure, mobility of amorphous materials decreases and they all become solids. A solid amorphous state is called ***glassy*** in other words.

***Polymers** are materials whose molecules are long chains composed of a large number of atoms or atomic groups connected by chemical bonds.* The distinctive feature of their chemical structure determines their special physical properties.

Polymers differ markedly from low-molecular materials in mechanical properties. Solids are known to be characterized by high strength at small reversible deformations. Liquids have the ability to unlimited deformation at their very low strength. Polymers are materials whose mechanical properties combine properties of solids and liquids; they are quite strong and, at the same time, can withstand quite large reversible deformations.

Polymeric materials include almost all natural (incl. plant) materials, such as wool, leather, horn, hair, silk, cotton, Indian rubber, etc., as well as all kinds of synthetic materials such as synthetic rubber, plastics, fibers, etc.

Most natural polymeric materials are protein materials; simple proteins are albumin, globulin; complex ones are casein, keratin, and collagen. Agar-agar contains up to 85% of carbohydrates, mainly polysaccharides, which are also polymers.

Polymers have other special properties in addition to special mechanical ones. For instance, their solutions have increased viscosity; above the solution, their solvent vapor elasticity is less, and their osmotic pressure is greater than it should be for ideal solutions. Polymers can swell in liquids very much.

The long-chain structure of polymer molecules contributes to formation of films and fibers.

Nowadays, polymers are more widely used as dielectrics.

The simplest organic polymer is polyethylene, whose polymer chain or macromolecule is composed of multiple reiterating monomeric links formed by joining molecules of ethylene:

$$n\cdot(CH_2{=}CH_2) \rightarrow [-CH_2-CH_2-]n.$$

Polyethylene is a representative of linear polymers. *Linear* polymers are those with macromolecules consisting of long one-dimensional chains (fig. 10.2, a: A is a monomeric link). A branched polymer, in addition to the main chain, has side branches, that is, side chains (fig. 10.2, b).

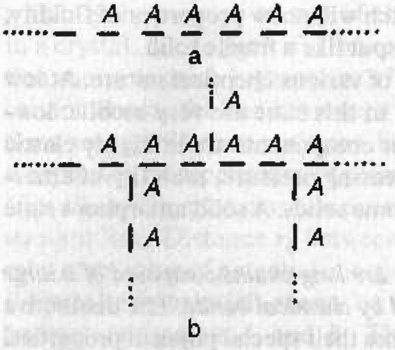


Fig. 10.2

Polymers composed of long chains connected one to another in a spatial network, are cross-linked or spatial ones, and those composed of similar polymers are homopolymers. Polymeric compounds, chains of which are composed of different monomeric links, are heteropolymers.

A polymer macromolecule is not rigid. Due to its thermal motion or under the influence of an external field, its spatial shape can change. These changes are called *conformational transformations.*

A freely jointed chain is extremely flexible (fig. 10.3). In such a chain, the valence angles are not fixed and rotation around them is free. In real polymeric chains, valence angles α have a certain value (fig. 10.4). This generates dependence of one chain link position on the position of the previous link. Such a chain has fewer conformations than a freely-jointed chain has, but the former is able to bend to a greater extent.

Because of their thermal motion, macromolecules have various conformations; on the one hand, the extremes are a rigid straight stick, on the other hand, an extremely flexible chain, curled into a ball (a globule).

Macromolecules can be huge with a relative molar mass from several thousands to hundreds of millions or even a billion. Due to the large dimensions of polymer molecules, its boiling point is extremely high (very high energy is needed to vaporize huge molecules). Whence, the decomposition temperature

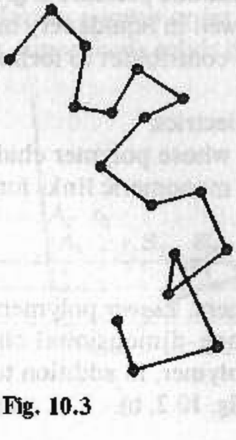


Fig. 10.3

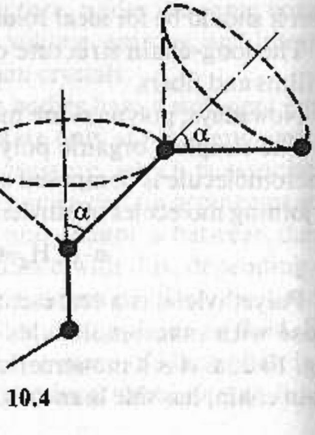


Fig. 10.4

of all polymers is lower than the boiling temperature and they are never in a gas state.

Therefore, polymers are in a condensed state, liquid or solid. Among solid polymers, *amorphous* and *crystalline* ones should be distinguished.

In a highly elastic state, an amorphous polymer can be deformed very much (up to 1000%); its deformation is reversible; irreversible flow is absent. In this sense, a highly elastic state is intermediate between liquid and solid. The highly elastic state of a polymer occurs due to flexibility of its macromolecules.

Macromolecules in all states of polymer are always more or less ordered, which leads to *supermolecular structures*. It is known that polymers are characterized by a large variety of supermolecular structures not only in crystalline, but also in their amorphous state. The primary elements of these structures are polymer molecules rolled up into globules or expanded into a linear macromolecule. When globules join one another, globular structures can be formed containing a large number of molecules, sometimes up to 1000. When expanded macromolecules join one another, elongated polymer bundles occur (fig. 10.5), which are of fluctuation nature: in place, they disappear, in another place they appear, but they also persist for quite a long time.

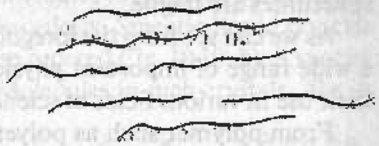


Fig. 10.5

The simplest primary supermolecular structures of bundles of polymer chains are present both in non-crystalline and crystalline polymers. During crystallization, bundles are folded up into "bands". In fig. 10.6, we can see bundles straightened (a) and folded up into a band (b). A trend to reduce surface tension generates folding of bands into plates (fig. 10.7) and formation of spherulites (fig. 10.8) or single crystals (fig. 10.9, single crystal of tobacco necrosis virus).

Numerous supermolecular structures were divided by the academician V.A. Kargin into four main types: ***globular*** (folded single molecules or groups of molecules), ***striped*** (structures of all polymers in highly elastic state),

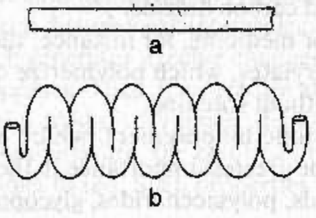


Fig. 10.6

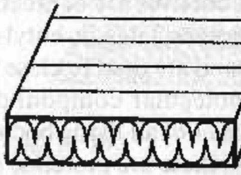


Fig. 10.7

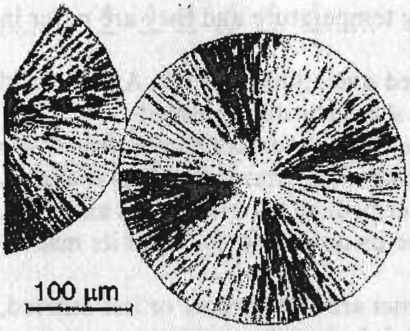


Fig. 10.8

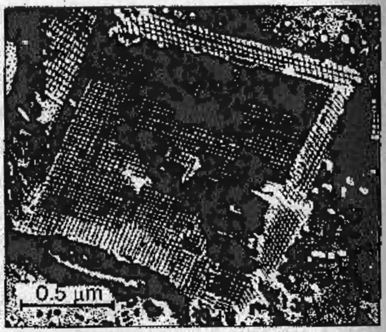


Fig. 10.9

fibrillar (linear bundles or their sets retaining their elongated shape), ***large-structural*** (spherulites, single crystals, etc.) ones.

Shapes and dimensions of supermolecular structures exert a great influence on the strength of polymers. For instance, a sample with small spherulites has high strength and good elastic properties, while samples with large spherulites are fragile.

As we can see from the foregoing, polymeric materials are characterized by a wide range of important physical and chemical properties, which permits their use in various fields of science and technology as well as in medicine.

From polymer such as polyethylene, polyvinyl chloride, etc., processed by pressure easily, a broad spectrum of medical tools and devices are produced.

Teflon, nylon and Dacron, Mylar, silastic polymer have a high chemical resistance, that is why they are used for production of prosthetic appliances for internal organs (blood vessels, heart valves, tendons, implantable eye lenses etc.). A polymer solution of polyvinylpyrrolidone is a good substitute of blood plasma.

Nowadays, cellophane membranes are used in an artificial kidney. Such membranes filter proteins and cellular components of blood. There are experiments to design artificial lungs with silicone membranes with a high transmission capacity with respect to oxygen and carbon dioxide.

Tissue adhesives are of great interest for medicine, for instance, there are alkyl-α-cyanocrylates, *n*-butyl-α-cyanocrylates, which polymerize quickly into a film and are used to close wounds without stitching.

High-molecular compounds also include biopolymers, which are the structural basis of all living bodies and of the greatest importance in their vital functions — these are proteins, nucleic acids, polysaccharides, glycoproteins, lipoproteins, glycolipids, etc.

§ 10.2. LIQUID CRYSTALS

***Liquid crystals** are materials that have properties of both liquids and crystals.*

In their mechanical properties, these materials are like liquids: they flow. With their optical properties, liquid crystals behave like anisotropic bodies (crystals): they are optically active, have double refraction, etc. Most often, liquid crystal properties become manifest in a certain temperature range above which it is an ***amorphous-liquid*** state, below this range, it is a ***solid crystal*** state.

Such duality of physical properties is due to the internal structure of liquid crystals. In them, mutual arrangement of molecules is intermediate between an amorphous state, in which there is no long-range order at all, and solid crystal state in which there is both long-range order of molecule center positions and an order in molecule orientation. Liquid crystal state is present in materials, whose molecules are elongated in the shape of a stick or plate. Such molecular shape generates possibility of their ordering.

According to the nature of molecular order, there are ***nematic and smectic liquid crystals.*** In nematic liquid crystals, molecules are oriented in parallel (fig. 10.10, a), but their centers are located disorderly. Smectic liquid crystals consist of parallel layers in which molecules are ordered (fig. 10.10, b). A special class consists of crystals of ***cholesteric***[1] type. Molecules in such crystals, like in smectic ones, are united in layers. However, inside each layer, the parallel arrangement of the axes of the molecules is like that of nematic state (fig. 10.10, b). There is also ordering between two layers: for passage to adjacent layers, general orientation of this layer changes with respect to general orientation of the previous layer by a small angle (the molecular structure is is twisted spirally).

The molecular structure of cholesteric liquid crystals is very sensitive to any slightest external influence. A small perturbation can disturb weak intermolecular forces which generates sizeable changes in optical properties. Thus, temperature affects the color of a crystal very much: depending on actual temperature, the color can be any — from violet to red. Such properties of liquid

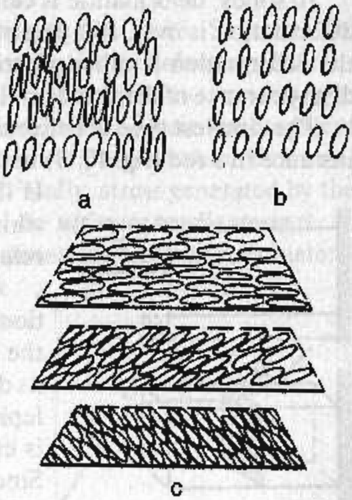


Fig. 10.10

---

[1] Their structure is typical for compounds containing cholesterol.

crystals are used for registering changes in temperature of different areas of bodies.

In medicine, this permits locating veins, arteries and other masses whose heat emission differs from that of adjacent media.

Liquid crystal materials are also used in various temperature-sensitive signaling devices.

The molecular structure of liquid crystals, and, hence, their optical properties, change in the presence of trace amounts of some chemical vapors. This fact makes it possible to use liquid crystals to detect traces of these substances.

Use of liquid crystals in devices and watches as digital displays is based on changes in their optical properties under the impact of electric field.

Studying liquid crystals in living bodies is a vast, little-studied, but extremely promising area.

§ 10.3. MECHANICAL PROPERTIES OF SOLIDS

Changes in mutual position of body points, which lead to a change in its shape and dimensions, are called ***deformation***.

Deformation can be caused by external impacts (mechanical, electrical or magnetic) or by changes in the body temperature. Here we consider deformations occurring from the effect of forces on a body.

In solids, deformation is called ***elastic*** if after termination of the force the deformation is over. If it persists after termination of an external influence, the deformation is called ***plastic***. The intermediate case, that is incomplete disappearance of deformation, is called ***elastoplastic*** deformation.

The simplest type of deformation is ***extension*** (***compression***). It occurs, for instance in a rod (fig. 10.11) under the action of a force directed along its axis. If the rod l long elongates by Δl, then $\varepsilon = \Delta l / l$, which is the measure of deformation and is called ***relative elongation***.

Another type of deformation is ***shear*** deformation (fig. 10.12). A force, which is tangent to one of the faces of a rectangular parallelepiped, causes its deformation, turning it into an oblique parallelepiped (see dash lines in the figure). Angle γ is called shear angle, and $\tan\gamma$ is relative shear. Since usually angle γ is small, we can further consider that $\tan\gamma = \gamma$.

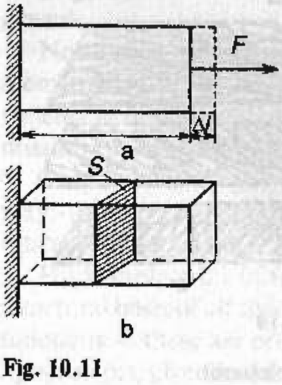


Fig. 10.11

Under the action of an external deforming force on a solid, the distance among atoms (ions) changes. This generates internal forces, trying to return atoms (ions) to their initial position. A measure of these forces is mechanical stress (or just stress).

Stress is not measured directly. In a number of cases, it can be calculated by external forces acting on the body. Stress can be indirectly determined by some physical effects (see, for example § 25.5).

As applied to extension deformation, ***stress*** σ can be expressed as the ratio of force to the cross-section area of the rod (fig. 10.11, b):

$$\sigma = F/S.$$

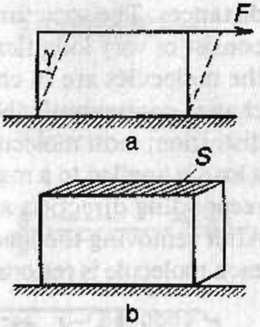


Fig. 10.12

For shear deformation, stress τ is expressed as the ratio of force to the area of facet, to which the force is tangential (fig. 10.12, b). In this case, τ is called tangential ***stress***:

$$\tau = F/S.$$

Small elastic deformations meet Hooke's law, according to which stress is proportional to deformation. For the two cases considered above (stretching, compression), this is analytically written as:

$$\sigma = E\varepsilon \quad \text{and} \quad \tau = G\gamma, \tag{10.1}$$

where E is Young's modulus; G is shear modulus.

An experimental extension curve is shown in fig. 10.13. Segment OA corresponds to elastic deformations, point B is the ***limit of elasticity*** characterizing maximum stress at which there are still no deformations remaining in a solid after termination of stress (residual deformations). Horizontal segment CD of the curve corresponds to ***flow stress***, which is stress at which deformation increases without any increase in stress. Finally, stress generated by the maximum load withstood before destruction is the ***ultimate tensile strength***.

Between elastic properties of crystalline monomers and polymer materials, there is a huge, fundamental difference: meeting its ***ultimate tensile strength***, steel is broken at the extension of 0.3%, but soft rubber can be stretched up to 300%. This phenomenon is associated with the fundamentally different mechanism of elasticity of high-molecular compounds.

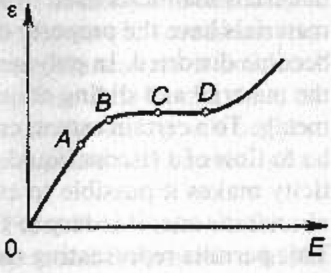


Fig. 10.13

As mentioned above, in deformation of crystal solids, such as steel, elastic forces are totally determined by changes in inter-atomic distances. The structure of high-molecular compounds is not regular. They consist of very long flexible molecules that are curved whimsically; parts of the molecules are in chaotic thermal motion so that their shape and length change continuously. Nevertheless, at any given moment, in a sample free of distortion, most molecules have a length close to the most probable one. When a load is applied to a material (fig. 10.14, a) its molecules straighten in the corresponding direction and the length of the sample increases (fig. 10.14, b). After removing the load, due to their chaotic thermal motion, the length of each molecule is restored, and the sample is shortened.

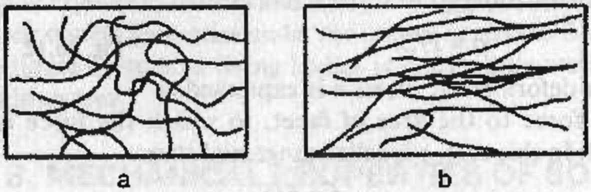


Fig. 10.14

Elasticity, which is inherent to polymers, is called ***rubber-like elasticity*** (high elasticity).

Here are data concerning mechanical properties of some materials (Table 10.1).

Table 10.1

Material	Young's modulus, GPa	Ultimate tensile strength, MPa
Steel	200	500
Glass filled nylon-6	8	150
Organic glass	3.5	50

The difference between deformation of crystal monomers and polymer materials manifests itself in its time dependence. The point is that almost all materials have the property of creep under the effect of a constant load, they become distorted. In polymers, straightening of molecules under the load of the material and sliding of macromolecules takes a longer time than creep in metals. To a certain extent, creep processes taking place in polymers are similar to flow of a viscous liquid. The combination of viscous flow and high elasticity makes it possible to call the deformation characteristic of polymers *viscoelastic* one. It is easy to simulate elastic and viscous properties of bodies. This permits representing mechanical properties of biological objects more clearly (see § 10.4).

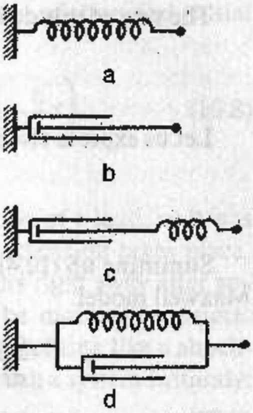


Fig. 10.15

As a model of an elastic body (elastic deformation), we choose a spring (fig. 10.15, a), whose small deformation meets Hooke's law.

The model of a viscous solid is a piston with holes moving in a cylinder with a viscous liquid (fig. 10.15, b).

The resistance force of the medium in this case will be taken proportional to the piston motion velocity [see (7.32)]:

$$F_{res} = r\frac{dx}{dt}. \tag{10.2}$$

Let us transform equation (10.2) based on an analogy. Instead of resistance force, we write down stress ($F_{res} \rightarrow \sigma$), that is a force divided by the unit of area; the friction coefficient, characterizing the property of a medium to resist a solid motion in it, we will replace by viscosity coefficient of the medium ($r \rightarrow \eta$), solid displacement will be replaced by its relative elongation ($x \rightarrow \varepsilon$). Then instead of (10.2), we get a relation between viscoelasticity rate and stress:

$$\sigma = \eta\frac{d\varepsilon}{dt}. \tag{10.3}$$

Validity of (10.3) can be partially verified by checking the dimensions: σ [Pa], η [Pa·s], $d\varepsilon/dt$ [s^{-1}]. From (10.3), it is clear that stress depends not on deformation itself, but its rate (velocity of the piston displacement).

Viscoelastic properties of bodies are simulated by systems consisting of different combinations of two simple models: "spring" and "piston". Let us consider some of them.

The simplest system, combining its elastic and viscous properties, is the Maxwell model, in which elastic and viscous elements are connected in series (fig. 10.15, c).

When a constant force acts, the spring is elongated instantly and elastically to the value determined by Hooke's law, and the piston moves with a constant velocity while the force (stress) is active. In this way, creep of a material is simulated.

If you stretch the Maxwell model quickly and, after that, fix this state, the deformation will persist. After its quick extension, the spring will begin to shrink, pulling the piston. In the course of time, there will be relaxation, that is, reduction (relaxation) of stress.

Let us describe this model mathematically. From Hooke's law, it follows that $\varepsilon_{el} = \sigma/E$, where ε_{el} is the elastic part of the total deformation in the Maxwell model.

The rate of this deformation is equal to

$$\frac{d\varepsilon_{el}}{dt}=\frac{1}{E}\cdot\frac{d\sigma}{dt}. \qquad (10.4)$$

Let us express viscous deformation rate from (10.3):

$$\frac{d\varepsilon_{visc}}{dt}=\frac{\sigma}{\eta}. \qquad (10.5)$$

Summing up (10.4) and (10.5), we find the rate of total deformation in the Maxwell model:

$$\frac{d\varepsilon}{dt}=\frac{d\varepsilon_{el}}{dt}+\frac{d\varepsilon_{visc}}{dt}=\frac{1}{E}\cdot\frac{d\sigma}{dt}+\frac{\sigma}{\eta}. \qquad (10.6)$$

From equation (10.6), it is possible to obtain time dependences of both deformation and stress.

If $\sigma = \text{const}$ and $d\sigma/dt = 0$ (a constant force is applied to the model), then from (10.6) it follows

$$\frac{d\varepsilon}{dt}=\frac{\sigma}{\eta} \quad \text{or} \quad d\varepsilon=\frac{\sigma}{\eta}dt.$$

Integrating the last expression from the initial point of time and zero deformation to the actual values of t and ε, we obtain

$$\int_0^{\varepsilon} d\varepsilon=\frac{\sigma}{\eta}\int_0^{\varepsilon} dt, \quad \varepsilon=\frac{\sigma}{\eta}t. \qquad (10.7)$$

This corresponds to creep (fig. 10.16, a).

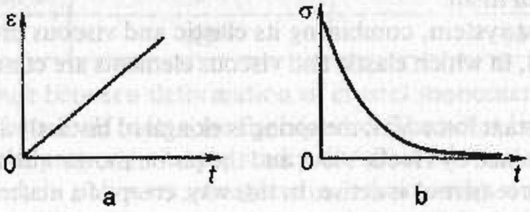


Fig. 10.16

If $\varepsilon = \text{const}$ and $d\varepsilon/dt = 0$ (constant deformation is maintained), then from (10.6), it follows

$$\frac{1}{E}\frac{d\sigma}{dt}=-\frac{\sigma}{\eta} \quad \text{or} \quad \frac{d\sigma}{\sigma}=-\frac{E}{\eta}dt.$$

Integrating the last expression from the initial point of time and initial stress σ_0 to the actual values of t and σ, we obtain

$$\int_{\sigma_0}^{\sigma} \frac{d\sigma}{\sigma} d\varepsilon = \frac{E}{\eta} \int_0^t dt, \quad \ln\frac{\sigma}{\sigma_0} = -\frac{E}{\eta} t, \quad \sigma = \sigma_0 \exp\left(-\frac{E}{\eta} t\right). \tag{10.8}$$

This corresponds to stress relaxation (fig. 10.16, b).

In the framework of the Maxwell model, under action of a load, as it was demonstrated, a quick (instantaneous) initial elastic extension takes place. In real polymers, viscoelastic deformation usually occurs right away after application of a load. Thus, Kelvin–Voigt model can be more appropriate: it consists of spring and piston connected in parallel, something like a shock-absorber in a car (fig. 10.15, d). If you generate stress in such a system instantly:

$$\sigma = \sigma_{el} + \sigma_{visc}, \tag{10.9}$$

applying a constant force, deformation of the system will be increasing. Using (10.1) and (10.3), we will transform (10.9):

$$\sigma = E\varepsilon + \eta \frac{d\varepsilon}{dt} \quad \text{or} \quad \frac{d\varepsilon}{\sigma - E\varepsilon} = \frac{dt}{\eta}.$$

Integrating the last expression from the initial point of time and zero deformation to the actual values of t and ε, we obtain

$$\int_0^{\varepsilon} \frac{d\varepsilon}{\sigma - E\varepsilon} = \frac{1}{\eta} \int_0^t dt, \quad -\frac{1}{E} \ln\frac{\sigma - E\varepsilon}{\sigma} = -\frac{t}{\eta},$$

$$\ln\left(1 - \frac{E}{\sigma}\varepsilon\right) = -\frac{E}{\eta}.$$

Finding the antilogarithm, we have

$$\varepsilon = \frac{\sigma}{E}\left(1 - \exp\left(-\frac{E}{\eta} t\right)\right).$$

As we can see, in the framework of the Kelvin–Voigt model, deformation increases in the course of time exponentially. Upon removal of the load ($\sigma = 0$ at point of time t_1), deformation will start to decrease exponentially. Both cases are shown in fig. 10.17.

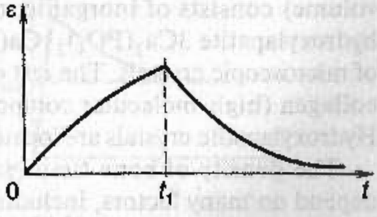


Fig. 10.17

In polymers, different types of deformation take place: elastic reversible deformation (a spring), viscoelastic reversible deformation (Kelvin–Voig model), and viscous irreversible deformation (a piston). A combination of these three elements makes it possible to create models that reflect mechanical properties of solids and, in particular, biological objects, most fully.

Modeling of mechanical properties of bodies is widely used in rheolog The main problem of rheology is to determine the dependence of stress on relative deformation: $\sigma = f(\varepsilon)$; dependence of stress on time (stress relaxation) — $\sigma = f(t)$ — at $\varepsilon = \text{const}$; dependence of relative deformation on time (creep) — $\varepsilon = f(t)$ — at $\sigma = \text{const}$.

§ 10.4. MECHANICAL PROPERTIES OF BIOLOGICAL TISSUES

Under mechanical properties of biological tissues, two varieties are understood. One is related to processes of biological mobility: muscle contraction in animals, cell growth, and motion of chromosomes in cells during their fission, etc. These processes are conditioned by chemical processes and supplied by ATP (adenosine tri-phosphate) energetically; but their nature is considered in the course of biochemistry. This specified group is conventionally called active mechanical properties of biological systems. The other variety is passive mechanical properties of biological systrems. Let us consider this problem in relation to biological tissues.

As a technical object, biological tissue is a composite material; it is formed by a spatial combination of chemically heterogeneous components. Mechanical properties of biological tissues differ from those of each component taken separately. Methods for determining mechanical properties of biological tissues are similar to methods of determining these properties in technical materials.

Bone tissue

Bone is the main material of the musculoskeletal system. In a simplified form, it can be considered that 2/3 of compact bone tissue mass (0.5 of its volume) consists of inorganic materials; the mineral material of bone is hydroxylapatite $3Ca_3(PO_4)_2 \cdot Ca(OH)_2$. This material is present in the form of microscopic crystals. The rest of bone consists of organic material, mainly collagen (high-molecular compound, fibrous protein with high elasticity). Hydroxylapatite crystals are located between collagen fibers (fibrils).

The density of bone tissue is 2,400 kg/m^3. Its mechanical properties depend on many factors, including age, individual growth conditions of the body and, surely, on the area of the body.

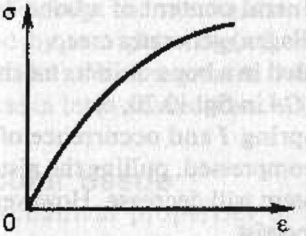


Fig. 10.18

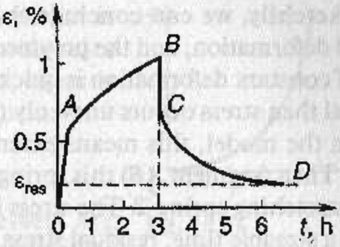


Fig. 10.19

The compositional structure of bone gives it the necessary mechanical properties: hardness, elasticity, and strength. Dependence $\sigma = f(\varepsilon)$ for compact bone tissue looks approximately like fig. 10.18, i.e., similar to the same dependence for a solid (see fig. 10.13); at small deformations, Hooke's law is valid. Young's absolute value is about 10 GPA, tensile strength is 100 MPa. It is useful to compare these data with the data for glass-filled nylon-6 (see Table 13, the correspondence is clear).

The approximate shape of creep curves of compact bone tissue can be seen in fig. 10.19. Segment *OA* corresponds to quick deformation, segment *AB* corresponds to creep. At time t_1, corresponding to point *B*, the load is removed. *BC* corresponds to quick deformation of contraction, *CD* corresponds to reverse creep. As a result, even in a long time interval, the bone sample does not restore its former dimensions, some residual deformation ε_{res} persists.

For this dependence, the following approximate model can be offered (fig. 10.20, a). The time dependence of relative deformation is shown in fig. 10.20, b. Under constant load, spring *1* is instantly stretched (segment *OA*), then, the piston is pulled out (relaxation segment *AB*), after termination of loading, there will be quick compression of spring *1* (*BC*), and spring *2* pulls the piston to its initial position (inverse relaxation *CD*). The model offered does not provide for residual deformation.

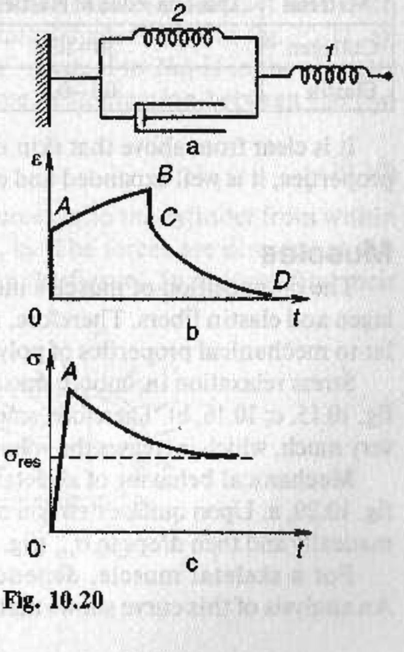


Fig. 10.20

Sketchily, we can conclude that the mineral content of a bone ensures rapid deformation, and the polymer part (collagen) generates creep.

If constant deformation is quickly generated in a bone or in its mechanical model then stress occurs unevenly (segment *OA* in fig. 10.20, c).

In the model, this means extension of spring *1* and occurrence of stress in it. Then (segment *AB*) this spring will be compressed, pulling the piston out and stretching spring *2*. The stress in the system will decrease. However, even after a sizeable time, residual stress σ_{res} will persist.

For the model, this means that at constant deformation, there will never arise a situation when the springs return to the non-deformed states.

Skin

It consists of collagen fibers, elastin (like, it is a fibrous protein) and the main tissue, which is the matrix. The collagen content is about 75% of the dry solid matter, and elastin content is about 4%. Approximate data of their mechanical properties are presented in Table 10.2.

Elastin stretches very much (up to 200–300%), almost like rubber does.

Collagen can stretch up to 10%, which corresponds to nylon fiber.

Table 10.2

Material	Absolute value of elasticity, MPa	Ultimate tensile strength, MPa
Collagen	10–100	100
Elastin	0.1–0.6	5

It is clear from above that skin is a viscoelastic material with highly elastic properties, it is well expanded and elongated.

Muscles

The composition of muscles includes connective tissue consisting of collagen and elastin fibers. Therefore, mechanical properties of muscles are similar to mechanical properties of polymers.

Stress relaxation in smooth muscles corresponds to the Maxwell model (see fig. 10.15, c; 10.16, b). Therefore, smooth muscles can stretch without much stress very much, which increases the volume of hollow organs, such as the bladder.

Mechanical behavior of skeletal muscle corresponds to a model shown in fig. 10.20, a. Upon quick extension of muscles by a certain value, stress rises dramatically and then drops to σ_{res} (fig. 10.20, b).

For a skeletal muscle, dependence $\sigma = f(\varepsilon)$ is non-linear (fig. 19.21). An analysis of this curve shows that approximately up to $\varepsilon = 25\%$, in the sartorius of a frog, the mechanism of deformation is caused by straightening of collagen molecules (see § 10.3). With a greater deformation, there is increase in inter-atom distances in molecules.

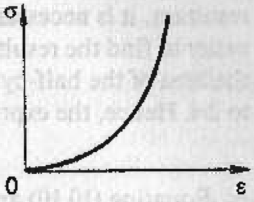


Fig. 10.21

Vascular tissue

Mechanical properties of blood vessels are mainly determined by properties of collagen, elastin and smooth muscle fibers. The content of these components of vascular tissue changes along the blood circulatory system: the ratio of elastin to collagen in the common carotid artery is 2:1, and in the femoral artery, the ratio is 1:2. The farther the distance from the heart, the greater the proportion of smooth muscle fibers is; in arterioles, they are the main component of vascular tissue.

In a detailed study of mechanical properties of vascular tissue it should be distinguished how the sample is cut out from the vessel (along or across the vessel). However, it is possible to consider deformation of a vessel as a whole because of action of pressure from within on an elastic cylinder.

Let us consider the cylindrical part of a blood vessel *l* long, with wall thickness *h* and radius of the inner part *r*. The longitudinal and cross sections are shown in fig. 10.22. Two halves of the cylindrical vessel interact with each other along the sections of the cylinder walls (shaded areas in fig. 10.22, a). The total area of this "interaction section" is equal to $2hl$. If in the vascular wall there is mechanical stress σ, the force of interaction between the two halves of the vessel is:

$$F = \sigma \cdot 2hl. \tag{10.10}$$

This force is equilibrated by pressure forces onto the cylinder from within (they are shown with arrows in fig. 10.22, b). The forces are directed at different angles to the horizontal plane (see in the figure). In order to find their

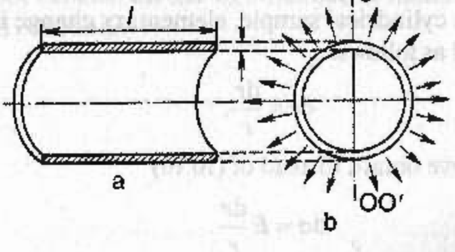


Fig. 10.22

resultant, it is necessary to sum up their horizontal projections. However, it is easier to find the resultant force if we multiply the pressure by the projection of the area of the half-cylinder onto vertical plane OO′. This projection is equal to $2rl$. Hence, the expression for force through the pressure has the form:

$$F = p \cdot 2rl. \tag{10.11}$$

Equating (10.10) and (10.11), we obtain $\sigma \cdot 2hl = p \cdot 2rl$, whence

$$\sigma = \frac{pr}{h}. \tag{10.12}$$

This is ***Lamé's equation.***

Let us assume that when the vessel is expanded, the volume of its wall does not change (the wall area increases and thickness decreases), i.e., the cross-sectional area of the vessel wall does not change (fig. 10.22, b):

$$2prh = \text{const}, \quad \text{i.e., } rh = b = \text{const}. \tag{10.13}$$

Taking into account (10.13), we transform (10.12):

$$\sigma = \frac{pr}{h} = \frac{prr}{rh} = \frac{pr^2}{b}. \tag{10.14}$$

From (10.14) it follows that in capillaries ($r \to 0$) stress is absent ($\sigma \to 0$).

Formula (10.14) describes the relation of three quantities, and so it is difficult to determine any dependence with this formula. Thus, dependence $\sigma = f(r)$ is not a dependence of $\sigma \sim r^2$, because the pressure and vessel radius are interdependent. Moreover, equation (10.14) does not include an absolute value of elasticity which is the basic mechanical characteristic of an elastic solid, so it is expedient to transform this formula. To do this, we differentiate (10.14) as a function of two variables:

$$d\sigma = \frac{r^2}{b}dp + \frac{2pr\,dr}{b}. \tag{10.15}$$

Having differentiated Hooke's law (10.1), we obtain:

$$d\sigma = E\,d\varepsilon. \tag{10.16}$$

As applied to a cylindrical sample, elementary change in relative deformation is represented as follows:

$$d\varepsilon = \frac{dr}{r}. \tag{10.17}$$

Using (10.17), we obtain instead of (10.16)

$$d\sigma = E\frac{dr}{r}. \tag{10.18}$$

Equating the right parts of (10.15) and (10.18), we find

$$E\frac{\mathrm{d}r}{r} = \frac{r^2}{b}\mathrm{d}p + \frac{2pr}{b}\mathrm{d}r. \tag{10.19}$$

Let us transform this equation:

$$\mathrm{d}p = \frac{b}{r^2}\left(E\frac{\mathrm{d}r}{r} - \frac{2pr\,\mathrm{d}r}{b}\right) = \left(\frac{Eb}{r^3} - \frac{2p}{r}\right)\mathrm{d}r. \tag{10.20}$$

If E is large, then from (10.20) it is possible to obtain approximately

$$\mathrm{d}p = \frac{Eb}{r^3}\mathrm{d}r. \tag{10.21}$$

Equations (10.20) and (10.21) can be used to find a relationship between pressure and the radius of a blood vessel, as well as its absolute value of elasticity. To solve the problem of pulse wave spreading, quantitative relations are also obtained on the basis of these equations.

In conclusion, we will enumerate sections and areas of medicine, for which it is particularly important to have an idea of passive mechanical properties of biological tissues:

- in space medicine, since a human is in new, extreme living conditions;
- sports achievements and their increase prompt sports physicians to pay attention to the physical capabilities of a human musculoskeletal system;
- hygienists protecting a human against vibrations should take in account mechanical properties of tissues;
- in prosthetics, replacing natural organs and tissues with artificial ones, it is also important to know mechanical properties and parameters of biological objects;
- in forensic medicine, it is necessary to know the stability of biological structures against various deformations;
- in traumatology and orthopedics, problems of mechanical impact on an body are decisive.

This list does not exhaust all the significance of material exposed in this chapter for medical education.

Chapter 11

Physical problems of hemodynamics

Hemodynamics is the field of biomechanics in which vascular blood circulation is studied. The physical basis of hemodynamics is hydrodynamics. Blood flow depends on both the properties of blood and those of blood vessels.

Physical principles of functioning of some technical devices associated with blood circulation are also discussed in this chapter.

§ 11.1. BLOOD CIRCULATION MODELS

Let us consider the hydrodynamic model of the circulatory system proposed by O. Frank.

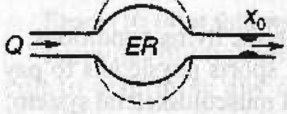


Fig. 11.1

In spite of being quite simple, it permits determining a connection between stroke volume (blood volume ejected by the ventricular contraction), hydraulic resistance of peripheral part of the circulatory system X_0 and a change in pressure in the arteries.

The arterial part of the circulatory system is simulated by an elastic reservoir (fig. 11.1, marked *ER*).

Since the blood is in the elastic reservoir, at any point of time its volume depends on pressure p meeting the following ratio:

$$V = V_0 + kp, \tag{11.1}$$

where k is the reservoir elasticity, the elasticity of the tank (pressure to volume ratio); V_0 is the volume of reservoir in absence of pressure ($p = 0$). Having differentiated (11.1), we obtain:

$$\frac{\mathrm{d}V}{\mathrm{d}t} = k\frac{\mathrm{d}p}{\mathrm{d}t}. \tag{11.2}$$

The elastic reservoir (arteries) receives blood from the heart; the volumetric blood flow is equal to Q. From the elastic reservoir, the blood flows away with volumetric blood flow Q_0 to the peripheral system (arterioles, capillaries).

Hydraulic resistance of the peripheral system is assumed to be constant. It is simulated by a "rigid" tube at the outlet of the elastic reservoir (see fig. 11.1).

It is possible to make an obvious equation (see fig. 11.1):

$$Q = \frac{dV}{dt} + Q_0, \tag{11.3}$$

showing that the volume rate of the blood flow from the heart is equal to the rate of increase in the volume of the elastic reservoir and the rate of blood outflow from the elastic reservoir.

Based on the Poiseuille equation (9.8) and formula (9.9), it is possible to write for the peripheral part of the system:

$$\frac{p - p_V}{X_0}, \tag{11.4}$$

where p is pressure in the elastic reservoir; p_V is venous pressure, it can be assumed to be zero; hence, instead of (11.4), we obtain:

$$Q = \frac{p}{X_0}. \tag{11.5}$$

Substituting (11.2) and (11.5) into (11.3), we obtain:

$$Q = k\frac{dp}{dt} + \frac{p}{X_0} \quad \text{or}$$
$$Q\,dt = k\,dp + \frac{p}{X_0}dt. \tag{11.6}$$

Let us integrate (11.6). The time limits of integration correspond to the cardiac cycle (cycle of the heart contraction) from 0 to T_p. These time limits correspond to equal pressures — the minimum diastolic pressure p_d:

$$\int_0^{T_p} Q\,dt = k\int_{p_d}^{p_d} dp + \frac{1}{X_0}\int_0^{p_d} p\,dt. \tag{11.7}$$

An integral with its equal lower and upper limits is zero, so from (11.7) we have:

$$\int_0^{T_p} Q\,dt = \frac{1}{X_0}\int_0^{T_p} p\,dt. \tag{11.8}$$

An experimental curve showing the time dependence of pressure in the carotid artery is shown in fig. 11.2 (unbroken line). The figure shows the cardiac cycle with systolic duration T_s and diastolic T_d, where p_s is the maximum systolic pressure.

The integral in the left-hand member of equation (11.8) is equal to the volume of blood ejected from the heart at one contraction: this is the stroke

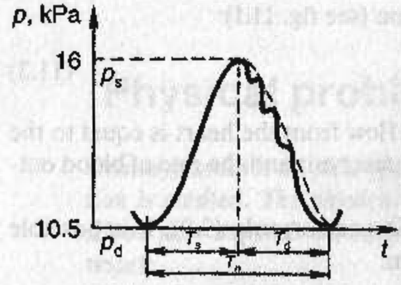


Fig. 11.2

volume. It can be found experimentally. In the right-hand member of equation (11.8), the integral corresponds to the area of the figure bounded by the curve and time axis (see fig. 11.2), this can also be found. Using the foregoing values of integrals, hydraulic resistance of the peripheral part of the circulatory system can be calculated by formula (11.8).

During a systole (heart contraction), the elastic reservoir stretches; after the systole (during diastole), blood flows away to the periphery, i.e. $Q = 0$. For this cycle, from (11.6) we have:

$$0 = k\,dp + \frac{p}{X_0}\,dt \quad \text{or} \quad \frac{dp}{p} = -\frac{dt}{kX_0}. \qquad (11.9)$$

Integrating (11.9), we obtain the time dependence of pressure in the reservoir after systole:

$$p = p_s \exp[-t/(kX_0)]. \qquad (11.10)$$

The dashed line in fig. 11.2 represents the corresponding curve. Based on (11.5) we obtain the time dependence of the blood outflow rate:

$$Q = Q_s \exp[-t/(kX_0)], \qquad (11.11)$$

where $Q_s = P_s/X_0$ is volumetric blood flow from the elastic reservoir at the end of systole (in the beginning of diastole).

Curves of dependencies (11.10) and (11.11) are exponents. Although this model describes real-life events quite approximately, it is extremely simple and represents the process at the end of diastole correctly. However, this model does not provide a description of pressure changes at the beginning of diastole.

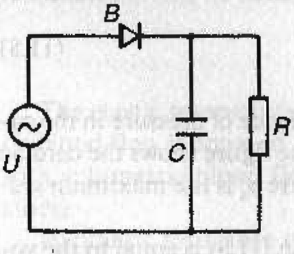


Fig. 11.3

Based on the mechanical model, by the analogy, an electric model can be drawn (fig. 11.3).

In this, source U generating electric power with alternating non-sinusoidal voltage is an analog of the heart, rectifier B simulates a heart valve. During the half-period, capacitor C accumulates a charge, and then discharges through resistor R, thus smoothing the current intensity passing through the resistor. Work of the capaci tor is similar to work of an elastic reservoir (aorta, arteries), which smoothes out fluctuations of the blood pressure in arterioles and capillaries. The resistor is an electrical analog of the peripheral vascular system.

In a more accurate model of the bloodstream, a greater number of elastic reservoirs is used knowing that the bloodstream is a spatial system. To take in account inertial properties of blood for creating the foregoing model, it was assumed that the elastic reservoirs, simulating the ascending and descending branches of the aorta have different elasticity. Fig. 11.4 shows the Roston model consisting of two reservoirs with different elasticity and inelastic links with different hydraulic resistances between reservoirs. The electric circuit shown in fig. 11.5 corresponds to this model. Here the current source generates energy with pulsating voltage $U(t)$ as an analog of pressure $p(t)$: capacitors C_1 and C_2 correspond to elasticity k_1 and k_2; electric resistances R_1, R_2 and R_3 correspond to hydraulic ones X_1, X_2 and X_3; current intensities I_1 and I_2 correspond to blood outflow rate Q_1 and Q_2.

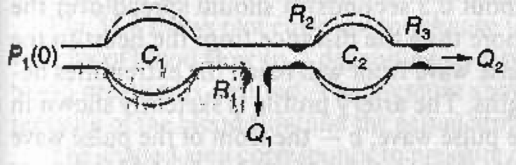


Fig. 11.4

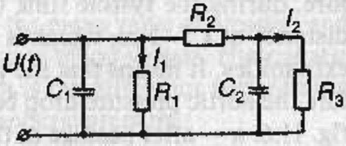


Fig. 11.5

This model is described by a system of two differential equations of the first order; their solution gives two curves corresponding to the first and second chambers.

The two-chamber model gives a better description of the processes taking place in the bloodstream; nevertheless, it does not explain pressure oscillations at the beginning of diastole.

Models that contain a few hundred elements are called models with distributed parameters.

§ 11.2. PULSE WAVE

Upon a contraction of the heart muscle (systole), blood is ejected from the heart into the aorta and arteries that diverge from it. If the walls of these vessels had been rigid, at the cardiac output, the blood pressure would have been transmitted to the periphery with the velocity of sound. Elasticity of vascular walls generates the following phenomena: during systole, the blood pushed out by the heart stretches the aorta, arteries, and arterioles, i.e., during systole,

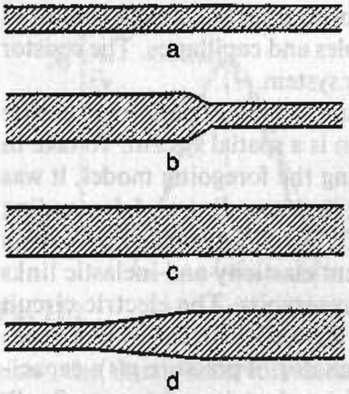


Fig. 11.6

large vessels receive a greater amoun of blood than that flowing away to the periphery. Human systolic pressure is normally about 16 kPa. During heart relaxation (diastole), the stretched blood vessel shrink and the potential energy transmitte to them by the heart through the blood transforms into kinetic energy of the blood flow; for all that, a diastolic pressure of about 11 kPa is maintained.

A wave of elevated pressure spreadin through the aorta and arteries, caused by ejection of blood from the left ventricl during the systolic cycle, is called ***pulse wave.***

The pulse wave spreads at velocity of 5–10 m/s and even more. Therefore, during the systole time (about 0.3 seconds) it should spread over the distance of 1.5–3 m, which is more than the distance from the heart to the extremities. It means that the pulse wave front will reach the extremities before the aortic pressure drop begins. The artery profile is sketchily shown in fig. 11.6: a — after passage of the pulse wave; b — the front of the pulse wave runs through the artery; c — pulse wave in the artery; d — high blood pressure drop begins.

Pulsation of blood flow will correspond to pulse wave velocity in large arteries; nevertheless, blood velocity (its maximum value is 0.3–0.5 m/s) is much less than the velocity of a pulse wave propagation.

As it is clear from the model experiment and general ideas of cardiac function, a pulse wave is not sinusoidal (harmonic). Like any periodic process, a pulse wave can be represented as a sum of harmonic waves (see § 7.4). So, we will consider a harmonic pulse wave as a model.

Let us suppose that a harmonic wave [see (7.45)] spreads along a blood vessel along X-axis with velocity υ. Blood viscosity and viscoelastic properties of vessel walls reduce the amplitude of the wave. We can consider (see § 7.5) that its attenuation will be exponential. Based on this, it is possible to write the following equation for the harmonic pulse wave:

$$p = p_0 \exp(-\chi^{x}) \cos\left[\omega\left(t - \frac{x}{\upsilon}\right)\right], \qquad (11.12)$$

where p_0 is pressure amplitude in the pulse wave; x is the distance to an arbitrary point from the source of oscillations (heart); t is time; ω is the circular frequency of oscillations; χ is a constant determining the wave attenuation. The pulse wavelength can be determined by the formula:

$$\lambda = \frac{\nu}{\upsilon} = \frac{2\pi\nu}{\omega}. \quad (11.13)$$

The pressure wave is a certain excessive pressure. Therefore, taking into account the main pressure of p_a (atmospheric pressure or pressure in the medium surrounding the vessel), the pressure change can be written as follows:

$$p = p_0 + p_0 \exp(-\chi^x)\cos\left[\omega\left(t - \frac{x}{\upsilon}\right)\right]. \quad (11.14)$$

As we can see from (11.14), as the blood is moving forward (as x is increasing) pressure oscillations are smoothed out. Fig. 11.7 gives a schematic representation of pressure oscillation in the aorta near the heart (a) and arterioles (b). The plot curves are assumed to meet the harmonic pulse wave model.

Fig. 11.8 shows plot curves of the change in the mean value of pressure and velocity of blood flow (υ_{bl}) depending on a type of blood vessels. Hydrostatic blood pressure is omitted. The pressure shown is the difference between the pressure of the blood flow and the actual atmospheric pressure.

The shaded area corresponds to pressure oscillation (pulse wave).

In large vessels, the pulse wave velocity depends on their parameters as follows (***Moens–Korteweg formula***):

$$\upsilon = \sqrt{\frac{Eh}{\rho d}}, \quad (11.15)$$

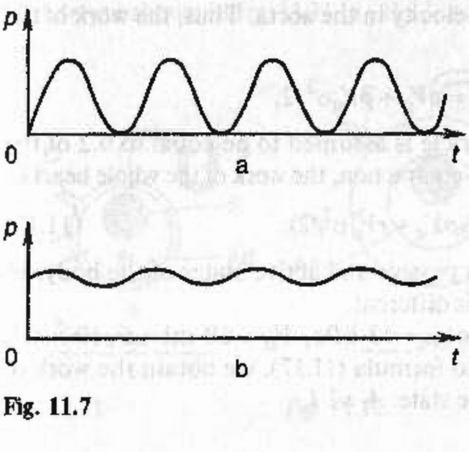


Fig. 11.7

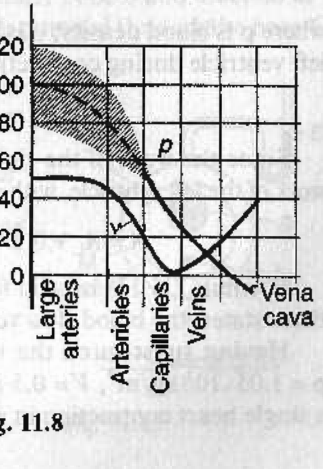


Fig. 11.8

where E is elasticity modulus; ρ is vessel material density; h is the thickness of vascular wall; d is vessel diameter.

It is interesting to compare (11.15) with the expression for the velocity of sound propagating in a thin rod:

$$\upsilon = \sqrt{\frac{E}{\rho}}. \qquad (11.16)$$

In humans, the absolute value of vascular elasticity increases with age; therefore, as it follows from (11.15), the pulse wave velocity also becomes higher.

§ 11.3. HEART WORK AND POWER. ARTIFICIAL BLOOD-CIRCULATION APPARATUS

The work done by the heart is spent to overcome the forces of pressure and imparting kinetic energy to the blood.

Let us calculate the work done with a single contraction of the left ventricle. Let us show V_{st} (stroke blood volume) as a cylinder (fig. 11.9). The heart is assumed to push this volume through the aorta with cross-section S at distance l at an average pressure p. At all that, the work being done is:

$$A_1 = Fl = pSl = pV_{st}.$$

To impart kinetic energy to this volume of blood the following work is spent:

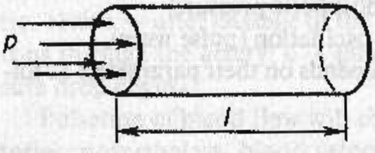


Fig. 11.9

$$A_2 = m\upsilon^2/2 = rV_{st}\upsilon^2/2,$$

where ρ is blood density; υ is blood velocity in the aorta. Thus, the work of the left ventricle during contraction is:

$$A_L = A_1 + A_2 = pV_{st} + pV_{st}\upsilon^2/2.$$

Since the work of the right ventricle is assumed to be equal to 0.2 of the work of the left ventricle, with a single contraction, the work of the whole heart is:

$$A = A_L + 0{,}2A_L = 1{,}2(pV_{st} + rV_{st}\upsilon^2/2). \qquad (11.17)$$

Formula (11.17) is valid for both passive and active states of the body. In these states, the blood flow velocity is different.

Having substituted the values of $p = 13$ kPa, $V_{st} = 60$ ml $= 6 \cdot 10^{-5}$ m^3, $\rho = 1.05 \cdot 10^3$ kg/m^3, $V = 0.5$ m/s into formula (11.17), we obtain the work of a single heart contraction in a passive state: $A_1 \approx 1$ J.

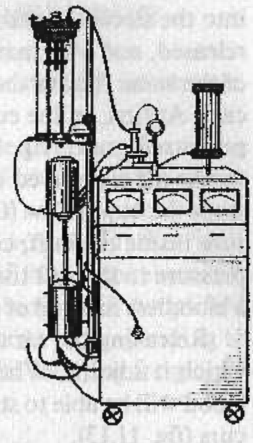


Fig. 11.10

Considering that on average the heart makes one contraction per second, we will find the heart work per day: $A_d = 86{,}400$ J. For active muscular activity, cardiac performance can increase several times.

If we take in account that duration t of a systole is about 0.3 s, the average power of heart during one contraction is $\langle W \rangle = A_1/t \doteq 3.3$ W.

In cardiac surgery, which requires temporarily cutting off the heart from the blood circulation system, a special artificial blood circulation apparatus is used (fig. 11.10). Essentially, this device is a combination of an artificial heart (a pump system) with artificial lungs (oxygenator is a system saturating the blood with oxygen).

§ 11.4. PHYSICAL PRINCIPLES OF THE CLINICAL METHOD OF BLOOD PRESSURE TAKING

A physical parameter such as blood pressure is of great importance in diagnostics of many diseases. In an artery, systolic and diastolic pressures can be taken using a needle connected with a pressure gauge directly. However, the bloodless method proposed by N.S. Korotkov is widely used in medicine. Let us consider the physical principles underlying this method on an example of taking blood pressure in the brachial artery.

A cuff is wrapped around the arm between the shoulder and elbow joint. A section of sleeve *M*, parts of the arm *Arm*, shoulder bone *S* and brachial artery *Art* are shown in fig. 11.11, a–11.13, a. As air is pumped through the hose *C*

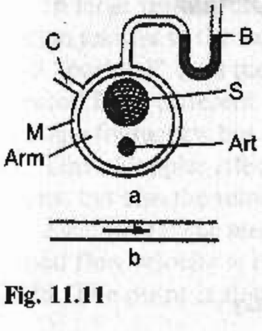


Fig. 11.11

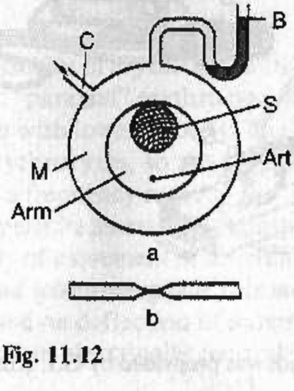


Fig. 11.12

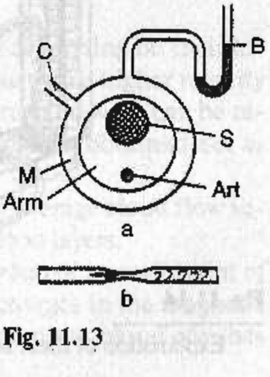


Fig. 11.13

into the sleeve, the arm is compressed. Then through the same tube, the air is released, and with manometer *B*, in the sleeve, air pressure is taken. Position b of the same figures shows longitudinal sections of the shoulder artery for each case. At first, in the cuff, the difference between air pressure and atmospheric pressure is zero (fig. 11.11), the cuff does not compress the arm and artery. As the air is pumped into the cuff, the latter squeezes the brachial artery and stops the blood flow (fig. 11.12). If the muscular system is relaxed, the air pressure inside the cuff, consisting of elastic walls, is approximately equal to the pressure in the soft tissues adjoining the cuff. This is the basic physical idea of a bloodless method of pressure taking.

Releasing the air they reduce the pressure in the cuff and in soft tissues, which it adjoins. When the pressure becomes equal to systolic pressure, the blood will be able to struggle through the artery squeezed: a turbulent flow occurs (fig. 11.13).

Characteristic tones and noises[1] accompanying this process are listened to by a physician taking the pressure; he or she places the phonendoscope onto a distal artery near the cuff (that is, at a greater distance from the heart). If we continue to reduce the pressure in the cuff, we can restore the laminar flow of the blood; which is manifested by dramatic reduction of the auscultated tones. The pressure in the cuff corresponding to restoration of laminar flow in the artery is recorded as diastolic pressure.

To take blood pressure, devices shown in fig. 11.14 are used: fig. a is a sphygmomanometer with a mercury manometer; fig. b is a sphygmomanometer with metal diaphragm manometer; *S* is cuff; *B* is a squeeze pressure pump; *P* is a manometer.

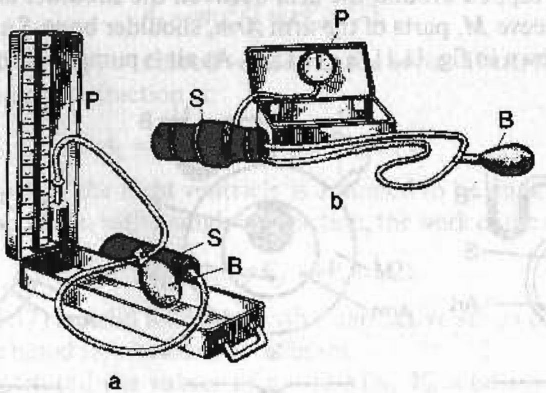


Fig. 11.14

---

[1] Explanation of these sounds was propvided by G.I. Kositsky.

§ 11.5. ASSESSING BLOOD FLOW VELOCITY

There are several methods of determining blood flow velocity. Let us consider the physical principles underlying two of them.

Ultrasonic method (ultrasonic flow metering) is based on Doppler effect (see § 7.11). From generator *1* of electrical oscillations of ultrasonic frequency (fig. 11.15), a signal is transmitted to ultrasonic emitter *2* and frequency comparison device *3*. Ultrasonic wave *4* penetrates blood vessel *5* and reflects from moving erythrocytes *6*. Reflected ultrasonic wave *7* falls onto receiver *8*, where it is converted into an electrical oscillation and amplified. Amplified electric oscillation gets into device *3*. Here, the oscillations corresponding to the incident and reflected waves, are compared, and the Doppler frequency shift in the form of an electric oscillation is selected:

$$U = U_0 \cos 2\pi\upsilon_d t.$$

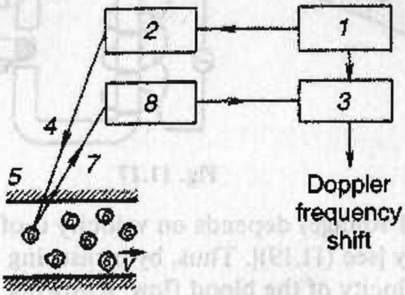


Fig. 11.15

By formula (7.62), the velocity of erythrocytes can be determined:

$$\upsilon_0 = \frac{\upsilon}{2}\frac{\nu_d}{\nu_G}. \qquad (11.18)$$

In large vessels, the velocity of erythrocytes varies depending on their location relative to the axis: "paraxial" erythrocytes move with a higher velocity and "parietal" ones move with lower velocity. The ultrasonic wave can be reflected from different erythrocytes, so the Doppler shift is obtained not as a single frequency, but as a frequency range.

Thus, Doppler effect permits assessing not only the average blood flow velocity, but also the velocity of movement of different blood layers.

Electromagnetic method (electromagnetic flux metering) of measurement of blood flow velocity is based on deflection of moving charges in the magnetic field. The point is that as an electrically neutral system the blood consists of positive and negative ions. Therefore, moving blood is a stream of charge particles that move with velocity υ_{bl}. On a moving electric charge q in an electric field with induction B, the following force acts (see § 16.3):

$$F = q\upsilon_{bl}B. \qquad (11.19)$$

If the charge is negative, the force is directed opposite to the vector produc of υ_{bl} B.

As shown in fig. 11.16, the forces acting from the magnetic field onto opposite charges are directed oppositely. Near one vascular wall, there is mor positive charge, near the other wall, in contrary, the charge is more negativ Redistribution of charges along the cross section of the vessel will generat an electric field. This physical phenomenon is called the ***Hall effect***.

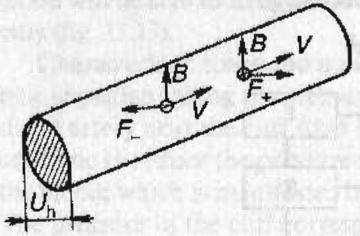


Fig. 11.16

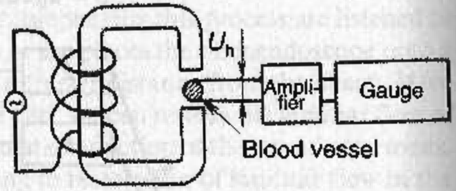


Fig. 11.17

Voltage U_h (Hall voltage) depends on velocity υ of ion movement, i.e., on the blood velocity [see (11.19)]. Thus, by measuring this voltage, one can to determine the velocity of the blood flow. Knowing cross-section area S of the vessel, it is not difficult to calculate the volumetric blood flow (m^3/s):

$$Q = \upsilon_{bl}S. \qquad (11.20)$$

It is more convenient practically to use a variable magnetic field in this method (fig. 11.17), which generates alternating Hall voltage U_h that is amplified and measured after that.

Section 3

EQUILIBRIUM AND NON-EQUILIBRIUM IN THERMODYNAMICS. DIFFUSION PROCESSES IN BIOLOGICAL MEMBRANES

The section deals with phenomena, the essence of which is determined by the chaotic motion of a huge number of molecules, which constitute bodies of different nature. Two main methods are used to study these phenomena. One of them is the thermodynamic method; it follows from basic experimental laws called the fundamentals (laws, principles) of thermodynamics. This approach does not take into account the internal structure of a material.

Another method is molecular kinetics (statistical method), which is based on the idea of molecular structure of a substance. By taking into account a large number of molecules in a substance, it is possible to determine certain laws using the theory of probability.

The section uses both approaches to varying extents.

It is important for physicians to understand energy properties of a body, heat exchange of biological systems with the environment, clarification of physical processes taking place in biological membranes, etc.

Chapter 12

Thermodynamics

Thermodynamics is a branch of physics that considers systems (thermodynamic systems), among which energy exchange is possible, without taking into account the microscopic structure of the bodies that make up the system. We distinguish thermodynamics of equilibrium systems, or systems passing to equilibrium (classical, or equilibrium thermodynamics, which is often called simply thermodynamics), and the thermodynamics of non-equilibrium systems (non-equilibrium thermodynamics). Non-equilibrium thermodynamics is of special importance when considering biological systems.

Along with thermodynamics, problems associated with use of low temperatures and heated media for therapeutic purposes, this chapter covers issues of thermometry and calorimetry too.

§ 12.1. BASIC CONCEPTS OF THERMODYNAMICS. FIRST LAW OF THERMODYNAMICS

The state of a thermodynamic system is characterized by physical quantities called *parameters* (volume, pressure, temperature, density, etc.).

If the parameters of a system do not change in the course of time upon interacting with the surrounding systems, the state of such a system is called *stationary.* Within a short time interval, examples of such states are the state of the interior of a working fridge, state of the human body, state of the air in a heated room, etc.

In different parts of the system, which is in a stationary state, the values of the parameters usually differ: the temperature in different parts of the human body, concentration of the diffusing molecules in different parts of the biological membranes, etc. Thus, in the system, constant gradients of certain parameters are maintained; chemical reactions can proceed at a constant rate.

The stationary state is maintained by flows of energy and materials passing through the system. The stationary state is schematically represented in fig. 12.1; the temperature is different at different points of the system (*1* and *2*). It is clear that a stationary state can involve such systems that either exchange both energy and material with surrounding systems (***open systems***), or at least exchange only energy (***closed systems***).

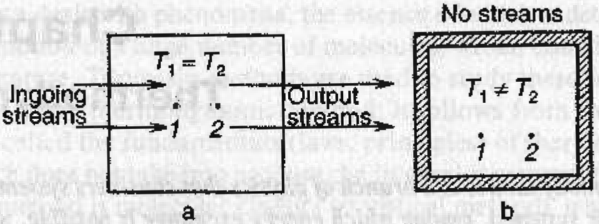


Fig. 12.1

A thermodynamic system that exchanges neither energy nor materials with surrounding bodies is called an *isolated system*. In the course of time, it comes to the state of thermodynamic equilibrium. Both in this state and in the stationary one, the system parameters remain unchanged over time. However, it is essential that in the equilibrium state, the parameters that do not depend on the mass or number of particles (pressure, temperature, etc.) are the same in different parts of this system.

Surely, no real thermodynamic system can be isolated, if only because it is impossible to surround it with a shell that does not conduct heat. An isolated system can be considered a convenient thermodynamic model. An equilibrium state of the isolated system is represented schematically in fig. 12.1, b.

Let us consider interaction of a closed system with surrounding bodies in more detail. Two different processes can be responsible for energy exchange between them: doing work and heat exchange.

The measure of energy transfer in the course of heat exchange is the amount of heat, and the same measure in the course of work is work[1].

Let us find an expression for a work done by gas when its volume changes. Let us assume that in a cylindrical vessel, under a piston, gas expands at a constant pressure from volume V_1 to volume V_2 (fig. 12.2), while the piston covers distance $\Delta l = l_2 - l_1$, and the volume changes by $\Delta V = V_2 - V_1$.

The piston with cross-sectional area S, at gas pressure p, is exposed to force $F = pS$. Since the direction of this force coincides with the direction of piston displacement, the work done by the gas is:

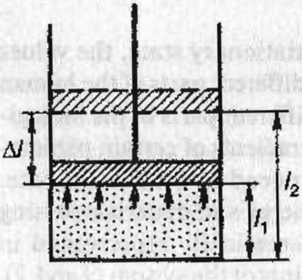


Fig. 12.2

$$A = F \cdot \Delta l = p \cdot S \cdot \Delta l = p \cdot \Delta V. \qquad (12.1)$$

---

[1] It is not quite good that "work" is applied to both one of possible processes of energy transfer and the measure of energy transfer in this process.

In case of gas expansion $\Delta V > 0$ the work is positive ($A > 0$); in case of gas compression $\Delta V < 0$ and $A < 0$. Note that this is in reference to work done by gas, but not by external forces. On the contrary, in case of the gas expansion, the work of all external forces will be negative, and in case of compression, it will be positive.

If gas pressure changes when the volume changes, it is necessary to calculate elementary work corresponding to a sufficiently small change in volume dV:

$$dA = p\,dV. \tag{12.2}$$

Integrating (12.2), we obtain the work done by gas:

$$A = \int_{V_1}^{V_2} p\,dV. \tag{12.3}$$

As an example, let us find work of ideal gas in isothermal process. To do this, instead of pressure, we substitute its expression from the Mendeleev–Clapeyron equation into formula (12.3):

$$p = \frac{m}{M}\frac{RT}{V}. \tag{12.4}$$

We obtain:

$$A = \int_{V_1}^{V_2} dV = \frac{m}{M}RT\int_{V_1}^{V_2}\frac{dV}{V} = \frac{m}{M}RT\ln\frac{V_2}{V_1}, \tag{12.5}$$

in this m is gas mass; M is molar mass (mole mass); T is thermodynamic temperature; $R = 8.31$ J/(mole · K) is the molar gas constant.

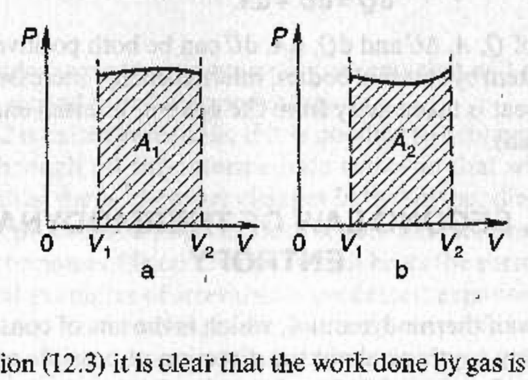


Fig. 12.3

From equation (12.3) it is clear that the work done by gas is graphically defined as an area of the curvilinear trapezoid in "pressure–volume" coordinates (fig. 12.3). From the figure, in which we can see plot curves of two different processes with the same initial and final states, it is clear that work depends on the process. Thus, work A_1 (fig. 12.3, a) is greater than work A_2 (fig. 12.3, b).

The law of energy conservation for thermal processes is formulated as the first law of thermodynamics.

The amount of heat transferred to a system is spent to change the internal energy of the system and performing work by the system:

$$Q = \Delta U + A. \tag{12.6}$$

The ***internal energy*** of the system is a sum of kinetic and potential energy of the particles that make up the system.

The internal energy is a function of the state of the system and has a definite value for the actual state; ΔU is the difference between two values of internal energy corresponding to the final and initial state of the system, $\Delta U = U_2 - U_1$.

Amount of heat, like work, is a function of the process, but not the state. Neither amount of heat nor work can be expressed as a difference between two values of any parameter in the final and initial state. In this connection, in (12.6), Q and A are written without the sign of increment Δ.

For sufficiently small values of Q, A and small increments of U, they use designations δQ, δA and $\mathrm{d}U$ respectively, by this emphasizing the difference of concepts of heat and work from the internal energy.

For simplification, the same designations will be further used ($\mathrm{d}Q$, $\mathrm{d}A$ and $\mathrm{d}U$); however, you should remember the difference between these physical quantities. Taking the above in account, the first law of thermodynamics can be written in the form of:

$$\mathrm{d}Q = \mathrm{d}U + \mathrm{d}A. \tag{12.7}$$

The values of Q, A, ΔU and $\mathrm{d}Q$, $\mathrm{d}A$, $\mathrm{d}U$ can be both positive (heat is transferred to the system by external bodies, internal energy increases, gas expands) and negative (heat is taken away from the system, internal energy decreases, gas is compressed).

§ 12.2. SECOND LAW OF THERMODYNAMICS. ENTROPY

The first law of thermodynamics, which is the law of conservation of energy, does not say anything about the direction of possible processes. Thus, according to the first law of thermodynamics, heat exchange is equally possible as spontaneous heat transfer from a more heated solid to a less heated body, and vice versa, from a less heated solid to a more heated body. However, from our everyday experience, it is well known that the latter process is unlikely in nature; so water cannot spontaneously be heated in a kettle due to cooling air in the room.

Another example: when a stone falls to the ground, it heats in proportion to the change in its potential energy; the reverse process— spontaneous elevation of the stone only because of its cooling — is impossible.

The second law of thermodynamics, as well as the first, is generalization of experimental data.

There are several formulations of the second law of thermodynamics: ***heat cannot pass from a solid with a lower temperature to a solid with a higher temperature by itself (Clausius formulation). In other words, a perpetual motion machine of the second kind is not possible (Thomson formulation), that is such a periodic process, whose sole result is conversion of heat to work due to cooling of one body.***

In a heat machine, the heat received from the heater does work, but at all that, a part of heat is necessarily transferred to a cooler. Fig. 12.4 gives a schematic representation of impossible (a) and possible (b) periodic processes according to the second law respectively.

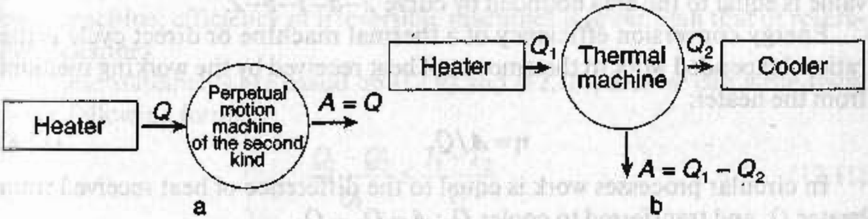


Fig. 12.4

Let us consider some thermodynamic concepts that will make it possible to quantify the second law of thermodynamics.

Process *1–2* is called ***reversible***, if it is possible to accomplish the inverse process *2–1* through all the intermediate states so that when the system returns to its initial states, there are changes in the surrounding bodies.

A reversible process is a physical abstraction. All real processes are irreversible if only because of friction force, which heats the surrounding bodies. Here are typical examples of irreversible processes: expansion of gas in the vacuum, diffusion, heat exchange, etc. To return a system to its initial state in all these cases, it is necessary external bodies must produce some work.

A cycle, or a circular process, is one in which a system returns to its original state.

The plot curve of a cycle is a closed line. The cycle shown in fig. 12.5, is ***direct***, it corresponds to a thermal machine, i.e. to a device that receives some amount of heat Q_1 from some solid — ***heat deliverer*** (heater), produces work and gives a part of this heat Q_2 to the other solid — ***heat receiver*** (cooler) (fig. 12.4).

In this cycle, the working medium (gas) does positive work (fig. 12.5). In process *1–a–2*, the gas expands, the work is positive and is equal numerically to the area under curve *1–a–2*. In process *2–b–1*, the work is negative (gas compression) and its absolute value is equal to the area under the corresponding curve. By algebraic summation, we obtain positive work in total; it is done by the gas in a cycle, and the absolute value of the work is numerically equal to the area bounded by closed curve *1–a–2–b–1*.

A reverse cycle[1] corresponds to work of a refrigeration machine, i.e. a system that takes heat from the cooler and transfers a greater amount of heat to the heater. As it follows from the second law of thermodynamics, this process (fig. 12.6) cannot run by itself, it takes place due to the work of external bodies. At all that, the gas does negative work: the work of compression in process *2–a–1* is negative, the work of expansion in process *1–b–2* is positive. By algebraic summation, we obtain negative work of gas, whose absolute value is equal to the area bounded by curve *2–a–1–b–2*.

Energy conversion efficiency of a thermal machine or direct cycle is the ratio of expended work to the amount of heat received by the working medium from the heater:

$$\eta = A/Q. \tag{12.8}$$

In circular processes work is equal to the difference of heat received from heater Q_1 and transferred to cooler Q_2: $A = Q_1 - Q_2$.

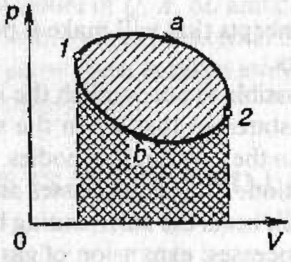


Fig. 12.5

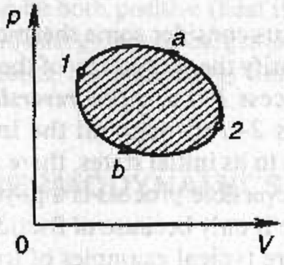


Fig. 12.6

---

[1] A reverse cycle should not be mixed up with the reversible one. The latter consists in a reversible process, it can be both direct (heat machine) and reverse (refrigeration machine).

Therefore,

$$\eta = (Q_1 - Q_2)/Q_1. \qquad (12.9)$$

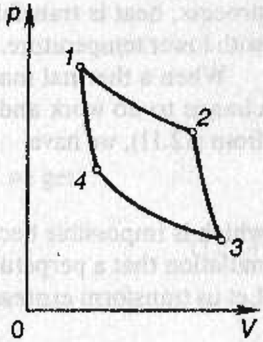


Fig. 12.7

Let us consider the Carnot cycle (fig. 12.7), consisting of two isotherms *1–2*, *3–4*, with corresponding temperature T_1 and T_2 ($T_1 > T_2$), and two adiabatic curves *2–3*, *4–1*. In this cycle, the working medium is an ideal gas. Transfer of an amount of heat from the heater to the working medium takes place at temperature T_1, and transfer from the working medium to the cooler takes place at temperature T_2. Without proof, we will mention that the efficiency of a reversible Carnot cycle depends only on temperature T_1 and T_2 of the heater and the cooler respectively:

$$\eta = (T_1 - T_2)/T_1, \qquad (12.10)$$

Carnot proved the following statements based on the second law of thermodynamics: all reversible machines working in a cycle consisting of two isotherms and two adiabatic curves with the same heater and cooler have equal efficiency which does not depend on the working medium or design of the cycle machine; efficiency of irreversible machines is lower than that of reversible machines.

These statements are based on (12.9) and (12.10), and we can write them in the following form:

$$\frac{Q_1 - Q_2'}{Q_1} \le \frac{T_1 - T_2}{T_1} \qquad (12.11)$$

or

$$1 - \frac{Q_2'}{Q_1} \le 1 - \frac{T_2}{T_1},$$

where sign "=" has relation to reversible cycle, and the sign "<" has relation to irreversible cycle.

This expression is a quantitative formulation of the second law of thermodynamics. Let us show that its consequence is both qualitative formulations given at the beginning of the section.

Let the first solid *receive* heat, and the second solid transfer it. Let us designate these heats Q_1 and Q_2'. Further, we will use the common term "received heat" and instead of transferred heat Q_2', we will write down received heat in the form of $Q_2 = Q_2'$. At all that, $Q_1 + Q_2 = 0$. Hence [see (12.11)] $T_1 - T_2 > 0$ and $T_1 > T_2$, which corresponds to Clausius' formulation: in a spontaneous process, heat is transferred from bodies with a higher temperature to those with lower temperature.

When a thermal machine expends all its energy received during heat exchange to do work and does not transfer energy to the cooler, $Q_2 = 0$, and from (12.11), we have:

$$(1-T_2/T_1)\geq 1,$$

which is impossible because T_1 and T_2 are positive. Whence, Thomson's formulation that a perpetual motion machine of the second kind is not possible. Let us transform expression (12.11):

$$1+\frac{Q_2}{Q_1}\leq 1-\frac{T_2}{T_1};\ \frac{Q_1}{T_1}+\frac{Q_2}{T_2}\leq 0. \qquad (12.12)$$

The ratio of the amount of heat, received or transferred by a working medium, to the temperature at which heat exchange takes place, is called reduced amount of heat.

Therefore (12.12), it can be formulated as follows: the algebraic sum of reduced amount of heat per cycle is not greater than zero (in reversible cycles, it is equal to zero, in irreversible cycles it is less than zero).

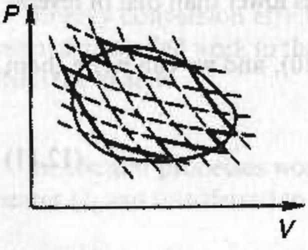


Fig. 12.8

If the state of a system changes not according to the Carnot cycle, but according to some arbitrary cycle, it can be represented as a set of sufficiently small Carnot cycles (fig. 12.8). Then expression (12.12) is transformed into a sum of sufficiently small-reduced amounts of heat, which is expressed in the limiting case by the following integral:

$$\oint\frac{dQ}{T}\leq 0. \qquad (12.13)$$

Expression (12.13) is valid for any irreversible ("<" sign) or reverse ("=" sign) cycle; dQ/T is elementary reduced heat. The circumference on the integral sign means that this is an integral about a closed contour, i.e. about a cycle.

Let us consider a reversible cycle (see fig. 12.5), consisting of two processes *a* and *b*. The following equation is valid for it:

$$\oint\frac{dQ}{T}\leq\underset{(a)}{\int_1^2\frac{dQ}{T}}+\underset{(b)}{\int_2^1\frac{dQ}{T}}.$$

Based on (12.13) for reversible cycles, we have

$$\underset{(a)}{\int_{1}^{2}}\frac{\mathrm{d}Q}{T}+\underset{(b)}{\int_{2}^{1}}\frac{\mathrm{d}Q}{T}=0.$$

By changing the limits of integration along path *b*, we get

$$\underset{(a)}{\int_{1}^{2}}\frac{\mathrm{d}Q}{T}+\underset{(b)}{\int_{1}^{2}}\frac{\mathrm{d}Q}{T}=0, \quad \text{or} \quad \underset{(a)}{\int_{1}^{2}}\frac{\mathrm{d}Q}{T}=\underset{(b)}{\int_{1}^{2}}\frac{\mathrm{d}Q}{T}. \qquad (12.14)$$

The latter means that the sum of reduced amounts of heat during reversible transition of the system from one state to another does not depend on the process, and for the given mass of gas is determined only by the initial and final state of the system. Fig. 12.9 shows plot curves of various reversible processes (a, b, c), for which initial *1* and final *2* states are common. The amounts of heat and work in these processes are different, but the sum of reduced amounts of heat is the same.

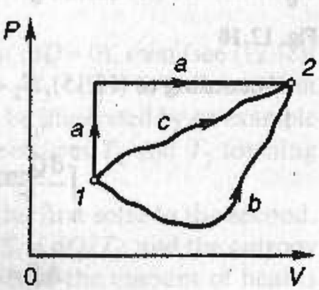


Fig. 12.9

A physical characteristic, independent on a process or displacement, is usually expressed as a difference between two values of some function corresponding to the final and initial state of the process or positions of the system. Thus, independence of gravity work from a trajectory makes it possible to express this work through the difference between potential energies at the end points of the trajectory. Similarly independence of an electrostatic field forces work from trajectories of the charge makes permits associating this work with the potential difference of the field points that are boundary ones when the charge moves. Similarly, the sum of reduced amounts of heat for a reversible process can be represented as difference between two values of some function of the system state, which is called *entropy*:

$$\Delta S = S_2 - S_1 = \int_{1}^{2}\frac{\mathrm{d}Q}{T}, \qquad (12.15)$$

where S_2 and S_1 is entropy in final *2* and initial *1* state respectively. Thus, entropy is a function of the state of a system, where the difference of values for two states is equal to the sum of reduced amounts of heat upon a reversible transition of the system from one state to another.

If the process is irreversible, then equation (12.15) is not correct. Let there be a cycle (fig. 12.10), consisting of reversible processes *2–b–1* and an irreversible one *1–a–2*.

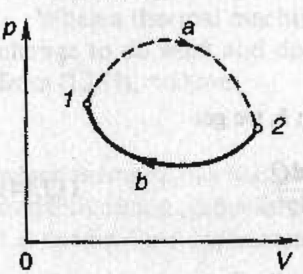


Fig. 12.10

Since a part of the cycle is irreversible, the whole cycle is irreversible, therefore, based on (12.13), we can write down:

$$\oint \frac{dQ}{T} \leqslant 0 \quad \text{or}$$

$$\int_1^2 \frac{dQ_{\text{irrev}}}{T} + \int_2^1 \frac{dQ_{\text{rev}}}{T} < 0. \qquad (12.16)$$

According to (12.15), $S_2 - S_1 = \int_2^1 \frac{dQ_{\text{rev}}}{T}$ then instead of (12.16) we obtain:

$$\int_1^2 \frac{dQ_{\text{irrev}}}{T} + S_2 - S_1 =< 0 \quad \text{or}$$

$$\Delta S = S_2 - S_1 > \int_1^2 \frac{dQ_{\text{irrev}}}{T}. \qquad (12.17)$$

Thus, in an irreversible process, the sum of reduced amounts of heat is less than the change in entropy. Adding right-hand parts of inequality (12.15) and (12.17), we obtain:

$$\Delta S \geq \int_1^2 \frac{dQ}{T}. \qquad (12.18)$$

where sign "=" is related to reversible processes and sign ">" is related to irreversible processes.

Relation (12.18) was derived from (12.11), and, therefore, also expresses the second law of thermodynamics.

Let us establish the physical meaning of entropy.

Formula (12.15) shows only entropy difference; entropy itself is determined within the accuracy of an arbitrary constant:

$$S = \int_1^2 \frac{dQ}{T} + S_0. \qquad (12.19)$$

If a system has passed from one state to another, then irrespective of the nature of the process — whether it is reversible or irreversible — its change in entropy is calculated by formula (12.15) for any reversible process taking place between these states. This phenomenon due to entropy being a function of the state of a system.

Entropy difference between two states is easily calculated for a reversible isothermal process:

$$S_2 - S_1 = \int_1^2 \frac{dQ}{T} = \frac{1}{T}\int_1^2 dQ = \frac{Q}{T},$$

where Q is the total amount of heat received by the system upon transition from state *1* to state *2* at constant temperature T. The last formula is used to calculate entropy in such processes as melting, vaporization, etc. In these cases, Q is the ***heat of phase transition***.

If the process takes place in an isolated system ($dQ = 0$), then [see (12.18)] in a reversible process, entropy does not change — $S_2 - S_1 = 0$, $S = \text{const}$, but in an irreversible process entropy increases. It can be illustrated by an example of heat transfer between two bodies with temperatures T_1 and T_2 forming an isolated system where $T_1 > T_2$ respectively.

If a small amount of heat dQ transfers from the first solid to the second, then the entropy of the first solid decreases by $dS_1 = dQ/T_1$, and the entropy of the second solid increases by $dS_2 = dQ/T_2$. Since the amount of heat is small, we can assume that the temperature of the first and second body does not change. **The total change in the entropy of a system is positive:**

$$dS = -dS_1 + dS_2 = \frac{dQ}{T_2} - \frac{dQ}{T_1} > 0,$$

therefore, the entropy of an isolated system increases. If in this system there had been spontaneous transition of heat from the solid with a lower temperature to a solid with a higher temperature, the entropy of the system would have decreased:

$$dS = dS_1 - dS_2 = \frac{dQ}{T_1} - \frac{dQ}{T_2} < 0,$$

but it contradicts (12.18). Thus, in an isolated system, such processes that result in a decrease in entropy of the system (the second law of thermodynamics) are impossible.

In an isolated system, an increase in entropy is finite. In the example above, the temperatures of the bodies will be equalized in the course of time, heat transfer between them will stop, and an equilibrium state will be achieved (see § 12.1). In this state, the parameters of the system will remain unchanged, and the entropy will reach its maximal value.

According to the molecular kinetic theory, entropy can be described best as a measure of disorder state of the system particles. Thus, as the volume of gas decreases, its molecules have to occupy more and more definite positions one relative to the other, which corresponds to a greater order in the system, for all that, entropy decreases. When a gas condenses or a liquid crystallizes at a constant temperature, heat is released, entropy decreases, and there is greater order in arrangement of particles.

A chaotic state of a system is characterized quantitatively by thermodynamic probability W_{ther}. To clarify this, let us consider a system consisting of four gas particles: *a*, *b*, *c*, *d* (fig. 12.11). These particles are in space, divided into two equal cells mentally, and can move in it freely.

The state of the system determined by the number of particles in the first and second cell, we call *macrostate*; a system state determined by which specific particles are in each cell, we call *microstate*. Then (fig. 12.11, a) the macrostate with one particle in the first cell and three particles in the second cell is implemented by four microstates. The macrostate corresponding to uniform arrangement of particles, two in each cell, is implemented by six microstates (fig. 12.11, b).

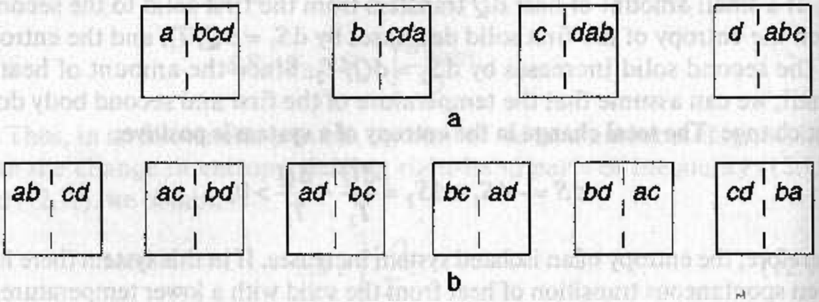


Fig. 12.11

Thermodynamic probability is the number of ways to place particles or the number of microstates that implement this macrostate.

In the above examples, $W_{\text{ther}} = 4$ in the first case, and $W_{\text{ther}} = 6$, in the second case. It is evident that a higher thermodynamic probability corresponds to uniform distribution of particles in the cells (two by two). At the same time, uniform distribution of particles corresponds to an equilibrium state with the highest entropy. According to the probability theory, it is clear that the system, left to its own devices, tends to the macrostate, which is implemented in the greatest number of ways, in the largest number of microstates, i.e., in the state with the highest thermodynamic probability.

Let us note that if a gas is given an opportunity to expand, its molecules will tend to occupy the entire volume uniformly; in this process, entropy increases.

A reverse process is the tendency of molecules to occupy only a part of a space, for instance, one half of the room, would be impossible — this would mean a state with a much lower thermodynamic probability and less entropy.

Whence, we can conclude that entropy is associated with thermodynamic probability. L. Boltzmann established that entropy is proportional to the logarithm of thermodynamic probability:

$$S = k \ln W_{\text{ther}}, \tag{12.20}$$

where k is the Boltzmann constant.

The second law of thermodynamics is a *statistical law* in contrast to Newton's first or second law.

The premise of the second law of thermodynamics that some processes are impossible is, in essence, the idea of extremely low probability, almost improbability, that is practical impossibility.

On the cosmic scale, there are sizeable deviations from the second law of thermodynamics, thus this law is not applicable to the entire Universe, nor to systems consisting of a small number of molecules.

In conclusion, we note again that **the first law of thermodynamics describes the energy balance of a process, but the second law of thermodynamics shows its possible direction**. Similarly the second law of thermodynamics complements the first law to a great extent; entropy complements the concept of energy.

§ 12.3. CRITICISM OF THE "HEAT DEATH" THEORY

This section was written during predominance of Marxist-Leninist philosophy. From the standpoint of modern physics, it is absolutely inadequate and is kept as an example absurdity resulting from mixingideology with science.

Clausius and later some other scientists transpolated the thesis of entropy growth of an isolated system to the entire Universe, confirming that entropy tends to its maximum value. When the maximum is attained, all energy types will convert to kinetic molecular energy; temperatures, concentrations etc. will become the same, biological processes will stop, and the Universe will experience heat death.

From its very creation, the heat death theory has been criticized by both natural scientists and philosophers. They have shown its inconsistency as a physical theory and as a philosophical concept.

First, conclusions obtained for an isolated system should not be transpolated to the entire Universe, because isolation itself presupposes finiteness and presence of some bodies beyond the system, while the Universe is infinite. Boltzmann refuted Clausius' conclusion about inevitable "heat death" of the Universe considering the following: since the Universe consists of a huge number of particles, in principle, there are cases when in some parts enormous fluctuations (heterogeneities) exist. Indeed, astrophysical data indicate that even nowadays, new stars are born from atomized matter, i.e. energy concentration processes take place. This proves that the second law of thermodynamics, as it is noted above, is not applicable to phenomena on the scale of the Universe.

According to modern ideas, homogeneous distribution of equal temperature materials is a very rare phenomenon and does not correspond to the maximum value of entropy, if we take into account the presence of gravitation.

The Universe is unsteady; it expands, under the influence of gravity, the heterogeneity of materials increases. There spring up galaxies, stars, and planets. In the presence of gravitation, this natural process is matched by increase in entropy.

From the "heat death" theory, the following question springs up: if the Universe will ever come to its "heat death", why did not it experience it in the past? From this, we can conclude that the world did not always exist, it had its beginning and will die some day. Then we can speak of some creator of the Universe, which leads directly to religious views. Therefore, it is not surprising that contemporary religious leaders support the idea of "heat death" of the Universe.

§ 12.4. THERMODYNAMIC POTENTIALS

Diverse thermodynamic problems solved by an analytical method, requir special functions, such as ***thermodynamic potentials***. Knowing the expressio of thermodynamic potentials through independent parameters of the system, we can calculate other parameters and other characteristics of thermodynami processes.

Let us consider some thermodynamic potentials.

Let us substitute the expression of work (12.2) and amount of heat in a reversible process ($\mathrm{d}Q = T\,\mathrm{d}S$) into the formula of the first law of thermodynamics (12.7):

$$\mathrm{d}U = T\mathrm{d}S - p\mathrm{d}V. \qquad (12.21)$$

This expression, combining the first and second laws of thermodynami is a complete differential of internal energy.

Let us write down a general expression of a total differential for our case:

$$dU = \left(\frac{\partial U}{\partial S}\right)_V dS + \left(\frac{\partial U}{\partial V}\right)_S dV.$$

Comparing (12.21) with this formula, we obtain:

$$dU = \left(\frac{\partial U}{\partial S}\right)_V dS + \left(\frac{\partial U}{\partial V}\right)_S dV. \qquad (12.22)$$

Thus, the partial derivative of internal energy by entropy is equal to temperature, and the derivative by volume, taken with its reverse sign, is equal to pressure. Internal energy itself is a thermodynamic potential.

Another thermodynamic potential is ***Helmholtz energy (free energy)***:

$$F = U - TS. \qquad (12.23)$$

Let us find the total differential of (12.23):

$$dF = dU - TdS - SdT. \qquad (12.24)$$

Using the expression for dU (see 12.21), let us transform (12.24):

$$dF = TdS - pdV - TdS - SdT = -SdT - pdV. \qquad (12.25)$$

Like in the previous case, taking in account that dF is the total differential of the function of variables T and V, we obtain:

$$-S = \left(\frac{\partial F}{\partial T}\right)_V, \quad p = -\left(\frac{\partial F}{\partial V}\right)_T. \qquad (12.26)$$

The physical meaning of F is clear from (12.25), in case of $T = \text{const}$ $dT = 0$ and $dF = -pdV = dA$, that is reduction of free energy is equal to a work done by the system in isothermal process. Since warm-blooded bodies maintain a constant temperature, their work is assumed to be done at the expense of reducing free energy.

Let us consider another thermodynamic potential (Gibbs energy), made up in the form:

$$G = F + pV = U - TS + pV. \qquad (12.27)$$

The differential of Gibbs energy is equal to:

$$dG = dU - TdS - SdT + pdV + Vdp. \qquad (12.28)$$

Taking into account (12.21), we transform (12.28):

$$dG = TdS - pdV - TdS - SdT + pdV + Vdp = -SdT + Vdp. \qquad (12.29)$$

Comparing (12.29) with the expression for the total differential, we have:

$$S = -\left(\frac{\partial G}{\partial T}\right)_p, \quad V = -\left(\frac{\partial G}{\partial p}\right)_T. \qquad (12.30)$$

Gibbs energy is used in entropy and volume calculations under isobaric-isothermal conditions.

Let the system come to its equilibrium state in an irreversible isobaric-isothermal process. In this case [see (12.17)]:

$$\mathrm{d}Q < T\mathrm{d}S, \qquad (12.31)$$

and for the Gibbs energy differential, instead of equation (12.29), the following inequation is obtained:

$$\mathrm{d}G < -S\mathrm{d}T + V\mathrm{d}p. \qquad (12.32)$$

Since $\mathrm{d}T = 0$ and $\mathrm{d}p = 0$, then from (12.32) we have

$$\mathrm{d}G < 0.$$

Such a decrease in Gibbs energy will occur until the equilibrium state has been achieved, and a change in Gibbs energy is zero ($\mathrm{d}G = 0$). Thus, in the isobaric-isothermal non-equilibrium process, Gibbs energy decreases and is minimum in the state of thermodynamic equilibrium.

Other thermodynamic potentials (U, F et al.) behave similarly in the state of thermodynamic equilibrium.

For instance, in an isothermal system with constant volume, Helmholtz energy decreases when approaching equilibrium. $F = F_{\min}$ is minimum in the state of equilibrium, i.e. if $\mathrm{d}T = 0$ and $\mathrm{d}V = 0$.

§ 12.5. SYSTEMS WITH VARIABLE NUMBER OF PARTICLES. CHEMICAL AND ELECTROCHEMICAL POTENTIALS

The above was applicable to a case when the amount of material in a system remained unchanged. However, thermodynamics also studies systems in which the number of particles varies. In this case, a change in the internal energy of the system is caused not only by heat exchange ($\mathrm{d}Q = T\mathrm{d}S$) and by performing work ($\mathrm{d}A = p\mathrm{d}V$), but also by changes in the number of particles in the system. Therefore, instead of equation (12.21), the following should be written:

$$\mathrm{d}U = T\mathrm{d}S - p\mathrm{d}V + \mu\mathrm{d}N. \qquad (12.33)$$

In this, $\mathrm{d}N$ is a change in the number of particles in the system, and μ is a coefficient called chemical potential. If into the differentials of Helmholtz energy and Gibbs energy we substitute the total differential of internal energy not in the form of (12.21), but in the same way as it is recorded in (12.33), we will obtain the corresponding equations:

$$\mathrm{d}F = -S\mathrm{d}T - p\mathrm{d}V + \mu\mathrm{d}N, \tag{12.34}$$

$$\mathrm{d}G = -S\mathrm{d}T - p\mathrm{d}V + \mu\mathrm{d}N. \tag{12.35}$$

From (12.33), (12.34), and (12.35), at constant pairs of appropriate parameters (S, V), (T, V) and (T, p), we obtain:

$$\mu = \left(\frac{\partial U}{\partial N}\right)_{S,V} = \left(\frac{\partial F}{\partial N}\right)_{T,V} = \left(\frac{\partial G}{\partial N}\right)_{T,p},$$

i.e., the chemical potential is equal to a change in the thermodynamic potential per one particle in the corresponding process.

A chemical potential is substituted into (12.33) and subsequent formulas for particles of the same type (molecules). If a thermodynamic system consists of k different types of molecules, then the expression for the total differential of thermodynamic potentials is added into the sum instead of one term:

$$\sum_{i=1}^{k} \mu_i \, dN_i, \tag{12.36}$$

where μ_i is the chemical potential for the i molecules, and $\mathrm{d}N_i$ is their number.

Note that the chemical potential is not a thermodynamic one, but, similar to pressure, temperature and entropy, it is called a system parameter. Like any parameter, it is a function of other parameters, for instance, $\mu = \mu(T, V, N)$.

For ideal gases and ideal solutions, that is systems in which the interaction of molecules of the *i-th* component among themselves is neglected, the following expression for the chemical potential is used:

$$\mu_i = \mu_{0i} + RT \ln \frac{N_i}{N}. \tag{12.37}$$

Formula (12.37) is written for a mole of a substance. In this N_i is the number of particles (moleoules) of the the *i-th* component; N is the total number of articles: $N = \sum_{i=1}^{k} N_i$; μ_{0i} is the chemical potential of the pure *i-th* component ($\mu_{0i} = \mu_i$ when $N_i = N$).

In practice, formula (12.37) is used for real systems too.

If particles are charged (ions, electrons), and the system is in an electric field, then instead of the chemical potential, electric potential is used ($\tilde{\mu}_i$):

$$\tilde{\mu}_i = \mu_i + Z_i F_\varphi, \qquad (12.38)$$

where μ_i, is the chemical potential of the *i-th* type of these particles in absence of an electric field [see (12.37)]; Z_i is the charge number of a particle (taking into account its sign); F is Faraday's constant; φ is an electric potential.

Based on formulas (12.37) and (12.38), we write down an expression for a change of electrochemical potential of the the *i-th* component for transition of the system from one state to another:

$$\Delta\tilde{\mu}_i = \mu_{0i} + RT\ln\frac{N_{2i}}{N_{1i}} + Z_iF(\varphi_2 - \varphi_1), \qquad (12.39)$$

where indices 1 and 2 refer to different states of the thermodynamic system.

An electrochemical potential, as well as a chemical one, has an energy meaning. In the light of this, let us explain the importance of each term in formula (12.39).

In (12.39), the first summand corresponds to a work to be done to cause chemical changes in one mole of a component during transition from one state to another.

We transform the second summand taking into account that at the volume given $N_{2i}/N_{1i} = c_{2i}/c_{1i}$, where c_{1i} and c_{2i} are molar concentrations of the solution corresponding to states 1 and 2 of the system:

$$RT\ln\frac{N_{2i}}{N_{1i}} = RT\ln\frac{c_{2i}}{c_{1i}} \qquad (12.40)$$

(the work of a single mole, which is done when the molar concentration changes from c_{1i} to c_{2i}).

The third summand corresponds to the work against the forces of an electric field when the state of one mole of ions change.

§ 12.6. STATIONARY STATE CONDITION. PRINCIPLE OF MINIMUM ENTROPY PRODUCTION

The above problems of thermodynamics are moslty applicable to equilibrium processes or to processes that lead to equilibrium states. Such limitations make it possible to explain a trend of thermodynamic processes in an isolated system.

Real processes and states in nature and technology are non-equilibrium, and many systems are open. These processes and systems are considered in non-equilibrium thermodynamics.

Just as in equilibrium thermodynamics, a special state is the state of equilibrium, therefore, in non-equilibrium thermodynamics stationary states are of special importance.

Despite the fact that in a stationary state, irreversible processes taking place in a system (diffusion, thermal conductivity, etc.) increase entropy, the total entropy of the system does not change. How can we understand this contradiction?

Let us represent a change in entropy ΔS of a system as a sum of two summands:

$$\Delta S = \Delta S_i + \Delta S_e, \tag{12.41}$$

where ΔS_i is the entropy change caused by irreversible processes in the system; ΔS_e is the entropy change caused by the system interaction with external bodies (flows passing through the system).

Irreversibility of the processes implies $\Delta S_i > 0$, the stationary state implies $\Delta S = 0$; therefore, $\Delta S_e = \Delta S - \Delta S_i < 0$. It means that in products (matter and energy) entering the system, entropy is less than entropy in products leaving the system.

In a steady state state, as it is already noted, entropy is maximum, and Gibbs energy is minimum. For stationary states, I. Prigogine also found the extreme value of a certain function, having formulated the principle of minimum entropy production: in the stationary state of the system, due to non-reversible processes, the rate of entropy production has minimum value under actual external conditions, which prevent the system from attaining the equilibrium state ($dS_i/dt > 0$ and is minimum).

According to the Prigogine principle, in a system with a stationary state internal non-equilibrium processes (diffusion, thermal conductivity, chemical reactions, etc.) run in such a way so that the incessant increase of entropy is minimum. This means that the system is not able to come out of its stationary state due to its internal irreversible processes. Thus, if due to small deviations (fluctuations), the system deviated from its stationary state, the trend of the internal processes to reduce dS_i/dt would return the system back to this state.

Let us note that all the above, including Prigogine principle, is true under given and invariable external conditions. In case of changes in external influence (flows entering and leaving the system), the system leaves one stationary state and transits to another if new external conditions are maintained over time.

In biological systems, examples of transient processes between stationary states are generation of nerve impulses, muscle contraction, etc.

In a stationary state, thermodynamic potentials of the system do not change. Let us use the condition of Gibbs energy constancy to determine a connection between voltage on a biological membrane and the concentration of ions on both sides of it.

If $\Delta G = 0$, then from (12.35), taking into account (12.36), at constant temperature and pressure ($\mathrm{d}T = 0$ and $\mathrm{d}p = 0$), it follows

$$\sum_{i=1}^{k} \mu_i \, \mathrm{d}N_i = 0, \tag{12.42}$$

or $\mu_1 \mathrm{d}N_1 + \mu_2 \mathrm{d}N_2 + \ldots = 0$.

In a particular case, if $\mathrm{d}N$ particles move from the part of the system with chemical potential μ_1 into the subsystem with chemical potential μ_2, then

$$\mu_1 \, \mathrm{d}N_1 + \mu_2 \, \mathrm{d}N_2 = 0$$

or $\mathrm{d}N_1 = \mathrm{d}N_2$, $\mu_1 = \mu_2$,

$$\Delta\mu = 0. \tag{12.43}$$

If ions are transferred in presence of an electric field, instead of condition (12.43), the equation of electrochemical potentials should be written:

$$\tilde{\mu}_1 = \tilde{\mu}_2, \quad \Delta\tilde{\mu} = 0. \tag{12.44}$$

Equations (12.43) and (12.44) can be considered as a condition of equilibrium and stationary states. Condition (12.44) can be applied only to such parts of the system, between which there are no impenetrable partitions. Thus, for a biological membrane, it is possible to write (12.44) if we are sure that specific ions pass through the membrane and that the system is in equilibrium or stationary state as a whole.

For instance, it is known that ions K^+ pass through a biological membrane and inside $[K^+]_i$ and outside of a cell $[K^+]_0$, their concentrations are different[1]. Therefore, it is possible to apply condition (12.44) to these ions in a stationary state. Let us do this using equation (12.38):

$$RT \ln \frac{[K^+]_i}{[K^+]_0} + F(\varphi_2 - \varphi_1) = 0, \tag{12.45}$$

since $Z = 1$. In this, φ_i and φ_0 are the potentials inside and outside the cells respectively. Denoting $\varphi_m = \varphi_i - \varphi_0$, from 12.45) we obtain **Nernst equation:**

$$\varphi_m = -\frac{RT}{F} \ln \frac{[K^+]_i}{[K^+]_0}, \tag{12.46}$$

where φ_m is a Nernst potential.

---

[1] In this, subscript *i* is derived from the word *inside*; it should not be mixed up with the designation of *the i-th* component.

It is known that in a cell of muscle fiber of a frog, the ratio of potassium ions concentrations is equal to $[K^+]_i/[K^+]_0 = 48$. Substituting this value into the Nernst equation (12.46), we obtain:

$$\varphi_m = -\frac{8.3 \cdot 300}{96\,500} \ln 48 = -98 \text{ mV}.$$

As you can see, the potential inside the cell is negative relative to the potential outside the cell. Thus a potential difference occurs on the membrane. These problems are discussed in more detail in Chapter 13.

§ 12.7. BODY AS AN OPEN SYSTEM

Initial development of thermodynamics was stimulated by the needs of Industrial production. At this stage (19th century), the main achievements were in formulation of laws, development of methods, cycles and thermodynamic potentials in relation to idealized processes, equilibrium and reversible ones.

Thermodynamics of biological systems was not developing at that time. A striking exception to this rule was the work of Mayer, who, by the color of venous blood of sailors working in a tropical climate, formulated, in essence, the applicability of the law of conservation of energy (the first law of thermodynamics) to living systems.

The first law of thermodynamics as the law of conservation of energy is so widely understood that its applications to biological systems are not considered here, especially since the course of normal physiology covers such topics as metabolism and energy exchange, nutrition and thermoregulation and § 27.5 analyzes radiative heat exchange of a human with the environment. It is more essential to consider some problems connected with the second law of thermodynamics (entropy) and biological systems.

Biological objects are open thermodynamic systems. They exchange energy and materials with the environment.

Speaking in general, a living body is a developing system that is not in a stationary state. Usually, however, in any not too long time interval, biological system state is considered as stationary.

Let us consider some problems with this assumption. For a body — steady state system — we can write down (see § 12.6) $dS = 0$, $S = \text{const}$, $dS_i > 0$, $dS_e < 0$. It means that higher entropy should be in excretion products, not in food. Entropy of the body/environment system increases, because the system is assumed to be isolated, however the entropy of the body remains constant. Entropy is a measure of randomness of a system (see § 12.2) so it can be concluded that the order of the body is maintained at the expense of reducing the order of the environment.

For some pathological states, entropy of a biological system can increase ($\mathrm{d}S > 0$). This is due to lack of stationarity, increase in randomness. For instance, in case of oncologic diseases, there is chaotic, disordered cell growth.

Formula (12.41) can be reduced to the following form:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{\mathrm{d}S_i}{\mathrm{d}t} + \frac{\mathrm{d}S_e}{\mathrm{d}t}.$$

For a stationary state ($S = \text{const}$, $\mathrm{d}S/\mathrm{d}t = 0$) the formula looks like:

$$\frac{\mathrm{d}S_i}{\mathrm{d}t} = -\frac{\mathrm{d}S_e}{\mathrm{d}t}. \qquad (12.47)$$

Based on (12.47) it is clear that at the normal state of the body, at the expense of internal processes, the rate of change of entropy is equal to the rate of change of negative entropy at the expense of exchanging materials and energy with the environment.

According to the ***Prigogine principle***, $\mathrm{d}S_i/\mathrm{d}t > 0$ and is minimum; hence $\left|-\frac{\mathrm{d}S_e}{\mathrm{d}t}\right|$ has the minimum value too [see (12.42)]. Whence, we can make a conclusion that the rate of environment entropy change is also minimum if the stationary state of the body remains the same.

The basis of functioning of living systems (cells, organs, body) is maintenance of their stationary state on condition that diffusion processes, biochemical reaction, osmotic phenomena, etc. occur in the body.

When external conditions change, processes in the body develop so that its state is not the previous stationary state.

It is possible to specify some thermodynamic criterion of bodies and biological structures adjusting to changes of external conditions (adaptation). If external conditions change (temperature increases or decreases, humidity changes, composition of the environmental air changes, etc.), but the body (a totality of cells) is able to maintain stationary states, the body adapts to the actual changes and exists. If the body is not able to maintain a stationary state upon changes of the environmental conditions and gets out of this state, this causes death of the body. In this case, the body has not been able to adapt, that is, it failed to attain a stationary state corresponding to the conditions changed quite quickly.

Finally: the arguments given in this section are based on a body as a a stationary system, which is not too far from the state of equilibrium. In this case the Prigogine principle is valid. Living organisms are far from the equilibrium state. Therefore, within the framework of the assumptions made, it is impossible to explain, in particular, cell proliferation and emergence of new structures. In a highly non-equilibrium system, the Prigogine–Glansdorff principle should be taken into account, according to which the rate of entropy production decreases.

In this area, non-equilibrium thermodynamics closes up with synergetics, nevertheless, consideration of such problems is beyond the course.

§ 12.8. THERMOMETRY AND CALORIMETRY

Accurate temperature taking is an integral part of scientific research and technical work, as well as medical diagnostics and biology.

The range of known temperatures is very wide. The lowest temperature nowadays attained is about $2 \cdot 10^{-5}$ K. The upper limit of attainable temperatures is not limited by anything. The greatest temperature attained in terrestrial conditions upon explosion of a hydrogen bomb and was about 10^8 K.

In stellar interior, according to spectroscopic data, the highest temperature can attain 10^9 K or greater.

The range of environmental temperatures, in which biological systems can exist for a long or short time, maintaining their ability to function, is much narrower. Finally, the temperature range of living bodies themselves for their active life, is not great at all — from 0 to 90 °C).

In a broad range, methods of obtaining and taking temperatures are quite different. ***The area of applied physics in which methods of temperature taking and related problems are studied is called thermometry.***

As it is already known, temperature proper cannot be measured directly. To determine it, one should need to set a temperature scale: choose a thermometric material and a physical property associated with temperature (thermometric property), to agree on the initial point of reference and the unit of temperature. For this purpose, two main temperatures (*reference points*), corresponding to the temperatures of phase transitions, such as melting ice, and boiling water under certain external conditions, are usually chosen. The part of the scale between these points is called *fundamental interval.* One of the reference points (0 °C is the temperature of ice melting) is taken as the reference starting point, and a fraction of the fundamental interval is taken as the unit of temperature. Thus, 1 °C is 0.01 of the fundamental interval.

Temperature scales differ by the thermometric property or material. It is possible to develop a huge number of scales, differing one from another to a great extent: no property depends on temperature strictly linearly and, moreover, is determined by the nature of material.

The disadvantage of all empirical scales is their dependence on properties of thermometric material. A scale based on the second law of thermodynamics and independent from both properties and substance, is called *thermodynamic temperature scale*. Its reference point is the temperature of the triple point of water equal to 273.16 K. It is determined using the Carnot cycle. Measuring the amount of heat Q_0 and Q_s in isothermal processes of this cycle at temperature T_0 of ice melting and T_s of water boiling respectively, it is possible to find

$$T_s/T_0 = Q_s/Q_0.$$

Similarly, for arbitrary temperature T

$$T/T_0 = Q/Q_0,$$

where Q is the amount of heat imparted to the system in the isothermal process at temperature T. Temperature established in this way is called *thermodynamic* temperature.

The unit of thermodynamic temperature is ***Kelvin*** (K) equal to 1/273.16 of the thermodynamic temperature of the triple point of water. Kelvin as a unit of a temperature interval equal to 1/273.16 of the thermodynamic temperature interval between 0 K and the triple point of water.

Any empirical scale is reduced to the thermodynamic one by means of introducing corrections that take into account dependence of the thermal property of a given material on its temperature.

As the temperature is determined by a value of a characteristic of the thermodynamic substance, temperature taking consists in measuring such physical properties as volume, pressure, electrical, mechanical, optical, magnetic properties, etc. A variety of methods of temperature taking is associated with a great number of thermometric materials and properties used in this process.

Thermometers are devices for temperature taking; they consist of a sensing element, in which the thermometric property is implemented, and a gauge (dilatometer, manometer, galvanometer, potentiometer, etc.). A prerequisite for temperature taking is thermal equilibrium between the sensing element and the body whose temperature is taken.

Depending on the measured temperature intervals, the most common thermometers are ***liquid, gas thermometers, resistance thermometer, thermocouple such as a thermometer, and pyrometers.***

In a liquid thermometer, the thermometric characteristic is volume, the sensing element is a reservoir with a liquid (usually it is mercury or alcohol). In pyrometers, the intensity of radiation is used as a thermometric property. The fundamental difference between pyrometers and other thermometers is that pyrometer sensors are not in direct contact with the body. Pyrometers can be used for measurement of any high temperatures.

For taking ultra-low temperatures, paramagnets serve as a thermometric material, and the property measured is dependence of their magnetization on temperature.

A mercury thermometer used in medicine shows the ***maximum temperature*** and is called maximum thermometer. This indicator is due to its structure: the mercury reservoir is separated from the graduated capillary by a hair-thin constriction, which prevents mercury from returning into the reservoir as the thermometer is cooling down (fig. 12.12). There are also ***minimum thermometers*** showing the lowest temperature observed over a long time interval.

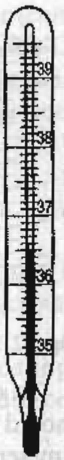


Fig. 12.12

A ***metastatic thermometer*** (fig. 12.13) is used to take temperature with high accuracy in a small value range; the device consists of large reservoir *1* with a liquid (usually it is mercury) and narrow long capillary *3.* The mass of mercury in reservoir *1* is variable; a part of it can be poured into reservoir *2,* so the lower limit of the measured temperature interval is set at scale mark 0. The scale sensitivity of such a thermometer is 0.01°. The measurement interval is only 5°, but it can be selected around different temperatures.

Many processes in physics, chemistry, and biology depend on temperature essentially, so obtaining and maintaining a certain temperature is an important problem. For this purpose, ***thermostats*** are used, in which constant temperature is maintained, which is carried out by automatic regulators or using the property of phase transitions to run at a constant temperature for this purpose.

To measure an amount of heat released or absorbed in various physical, chemical and biological processes, a number of methods are used, the totality of which is called ***calorimetry***[1].

Calorimetry methods are used to measure the heat capacity of bodies, the heat of phase transitions, dissolution, wetting, adsorption, heat accompanying chemical reactions, energy of radiation and radioactive decay, etc.

Such measurements are made using *calorimeters.* These instruments can be divided into two main types: calorimeters in which the amount of heat is determined by a change in their temperature, and calorimeters in which the temperature is constant and the amount of heat is determined by the amount of matter that has passed into another phase state (e.g., a melting solid).

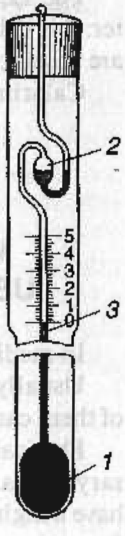


Fig. 12.13

---

[1] A group of methods for measuring thermal effects accompanying processes of vital functions are also called biocalorimetry, and the corresponding devices are called biocalorimeters.

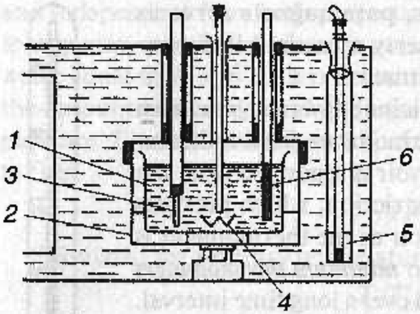


Fig. 12.14

Most calorimeters practically used are of the first type. In these cases, the amount of heat received by the calorimeter — body system can be found by a formula:

$$Q = c_c \Delta T.$$

where c_c is the heat capacity of the calorimetric system; ΔT is a change in its temperature, which would be there if there was no heat exchange with surrounding bodies.

To determine ΔT, the temperature change taken experimentally should should be adjusted taking into account the heat exchange between the calorimeter and environment. In this respect, all calorimeters can be divided into those with isothermal and adiabatic shell. To maintain isothermal or adiabatic conditions, the calorimeter has temperature regulators, as which contact thermometers are mostly used. Moreover, these are resistance thermometers and differential thermocouples.

Fig. 12.14 provides a schematic representation of a simple liquid calorimeter: *1* is calorimetric vessel; *2* is cylindrical vessel-shell; *3* is a heater; *4* and *5* are mixers; *6* is the thermometer.

Calorimeters can serve as thermostats too.

§ 12.9. PHYSICAL PROPERTIES OF HEATED AND COLD MEDIA USED FOR THERAPEUTIC PROPERTIES. USE OF LOW TEMPERATURES IN MEDICINE

In medicine, heated or cold bodies are used for local heating or cooling.

Usually, relatively accessible media are chosen for this purpose, some of them can exert a useful mechanical or chemical effect.

Physical properties of such media are due to their purpose. First, it is necessary that a long-term effect is produced. Therefore, the media used should have a high specific heat capacity (water, mud) or specific heat of phase transformation (paraffin, ice). Secondly, the media, applied directly onto a skin, should not cause pain. This, on the one hand, limits the temperature of such media, and on the other hand, prompts choosing media with a low heat capacity. Thus, water used for therapeutic purposes, has a temperature up to 45 °C, and peat and mud have the temperature of up to 50 °C as heat transfer (convection) in these media is less than in water. Paraffin is heated to 60–70 °C, as it has low thermal conductivity, and parts of paraffin, touching the skin directly, get cool quickly, crystallize and prevent heat flow from passing from its other parts.

Ice is used as a cooling medium for therapeutic purposes.

In recent years, low temperatures therapy is used in medicine on a quite large scale.

At a low temperature, preservation of separate organs and tissues is achieved out for the peurpose of transplantation, ensuring that vital activity and normal function is kept for a quite long time.

Cryogenic[1] *method* of destruction of tissue by freezing and defrosting, is used by physicians to remove tonsils, warts, etc. Special cryogenic devices and cryoprobes are developed for this purpose.

Using cold, which acts as an anesthetic, it is possible to destroy cerebral nuclear cells responsible for diseases, such as Parkinsonism.

In microsurgery, freezing ("adhesion") of wet tissues to a cold metal tool is used to capture and transfer these tissues.

New terms have sprung up in connection with medical applications of low temperature media: "*cryogenic medicine*", "*cryotherapy*", "*cryosurgery*", etc.

---

[1] *Cryo*... is a part of compound words meaning association with ice, low temperatures; cryogenic means associated with low temperatures.

Chapter 13

Physical processes in biological membranes

Biological membranes are an important part of a cell. They delimit a cell from the environment, protect it from harmful external influences, control the exchange between the cell and its environment, contribute to the generation of electrical potentials, participate in synthesis of universal ATP energy accumulators in mitochondria, etc. Essentially, membranes form the structure of a cell and carry out its functions. Many diseases (atherosclerosis, poisoning, etc.) are associated with a disorder of the structure and function of membranes.

This chapter discusses physical properties of biological membranes and basic physical processes that take place in them.

§ 13.1. STRUCTURE AND MODELS OF MEMBRANES

Membranes surround all cells (plasmatic or external cell membranes). Without a membrane, the contents of a cell would be simply spilt; diffusion would generate thermodynamic equilibrium, which means absence of life. We can say that the first cell sprang up when it was able to isolate itself from the environment with a membrane.

Intracellular membranes divide a cell into a number of closed compartments; each of them performs a certain function.

The thickness of membranes is about several nanometers, so it is invisible under an optical microscope (see § 26.8), but is visible under an electronic microscope (see § 28.2).

The basis of any membrane structure is a double lipid layer (mostly phospholipids). Molecules of lipids forming the membrane are aliphatic compounds, i.e., they consist of two functionally different parts: polar head and non-polar hydrophobic tail (fig. 13.1). The double lipid layer is formed from two monolayers of lipids so that the hydrophobic tails of both layers are directed inwardly. At the same time, the smallest possible contact of hydrophobic parts of molecules with water is ensured (fig. 13.2).

However, this understanding of the membrane structure failed to explain the arrangement of protein in the membrane (in some membranes, it is more

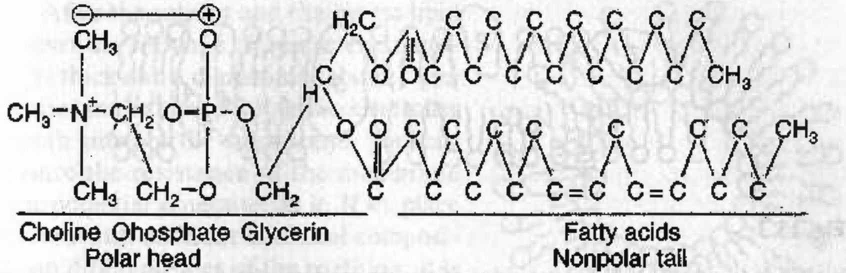


Fig. 13.1

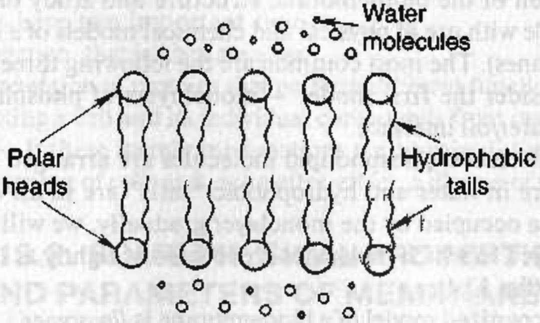


Fig. 13.2

than a half of the total mass) nor the question of membrane permeability for hydrophilic particles.

Later, many other hypotheses about the structure of biological membranes were put forward, but none became generally accepted. Nowadays, the most recognized idea is the one offered in 1972 by Singer and Nicolson — liquid-mosaic model, which is based on the same lipid bilayer membrane. This phospholipid base is a sort of two-dimensional solvent in which proteins are more or less submerged. Thanks to these proteins, specific functions of membranes, which are permeability, active transfer through the membrane, generation of electric potential, etc. are fully or partially implemented. The liquid-mosaic structure of a membrane is schematically represented in fig. 13.3. In this, *1* are surface proteins, *2* are half-submerged proteins, *3* are completely submerged (integral) proteins, *4* are proteins forming ion canal *5*.

Membranes are not fixed, "quiet" structures. Lipids and proteins mutually exchange their places and move both along the plane of the membrane (lateral diffusion) and across it (so-called "flip-flop"). High mobility of lipids corresponds to lateral diffusion, and low mobility, i.e. mutual exchange of places

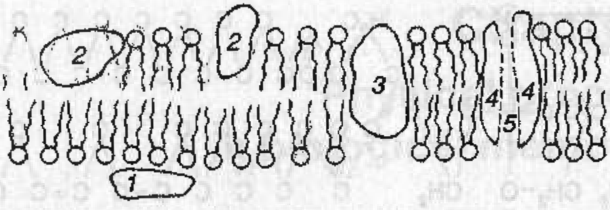


Fig. 13.3

by lipids located on different sides of the membrane, being a rare process, corresponds to "flip-flop".

Clarification of the biomembrane structure and study of its properties became possible with use of physical and chemical models of a membrane (artificial membranes). The most common are the following three models.

Let us consider the first model — monolayers of phospholipids on the water/air, or water/oil interface.

On such interfaces, phospholipid molecules are arranged so that hydrophilic heads are in water and hydrophobic "tails" are in air or in oil. If we reduce the area occupied by the monolayer gradually, we will finally end up with a monolayer, in which molecules are located as tightly as in one of membrane bilayers (fig. 13.4).

Another recognized model of a biomembrane is *liposomes.*

These are tiny bubbles (vesicles) consisting of bilipid membrane and obtained by processing a mixture of water and phospholipids with ultrasound. Liposomes are a sort of biological membrane, without any protein molecules. Liposomes are often used in experiments studying the effect of various factors, such as the composition of phospholipids, on membrane properties, or vice versa studying the effect of membrane environment on the properties of embedded proteins. Liposomes are shown schematically in fig. 13.5.

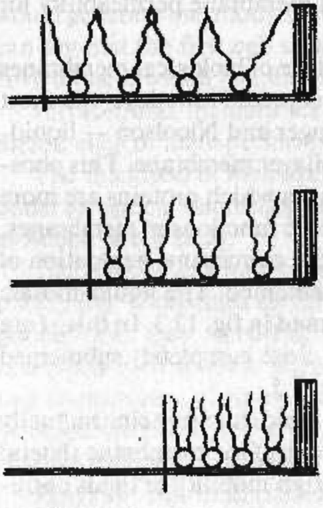


Fig. 13.4

The third model, which makes it possible to study some properties of biomembranes by direct methods, was ***bilipid*** (lipid bilayer) ***membrane*** (BLM).

P. Muller and his colleagues developed such a model membrane first in 1962. They filled a hole in a Teflon partition separating two water solutions, with phospholipid dissolved in heptane. After the solvent and the excess lipid spilt over the Teflon, a bilayer several nanometers thick and a diameter of about 1 mm is formed in the hole. Placing two electrodes on both sides of the membrane, you can measure the resistance of the membrane or the potential generated on it. If we place solutions with different chemical composition on different sides of the partition, it is possible to study permeability of the membrane for different agents.

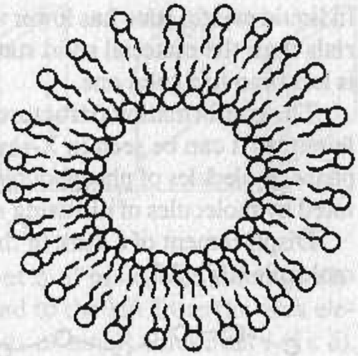


Fig. 13.5

Membranes have two important functions: matrix function, that is, they are a matrix, a base for retention of proteins that perform different functions, and barrier function protecting a cell and its individual compounds from invasion by undesirable particles. If these membrane functions are impaired, there are changes in normal functioning of cells and, as an after-effect, a disease of the body.

§ 13.2. SOME PHYSICAL PROPERTIES AND PARAMETERS OF MEMBRANES

Measurement of mobility of a membrane molecules and diffusion of its particles through the membrane evidences that the bilipid layer behaves like a liquid. At the same time, a membrane is an ordered structure. These two factors make us think that in the membrane, during its natural functioning, phospholipids are in a liquid crystal state (see § 10.2). Liquid properties of the membrane are confirmed by EPR (see § 30.3) and NMR (see § 30.4) methods.

The viscosity of the lipid layer of membranes is about two orders higher than the viscosity of water, it is equal to 30–100 MPa · s, which approximately corresponds to the viscosity of vegetable oil. Its surface tension is 2–3 orders lower (0.03–1 mN/m) than that of water.

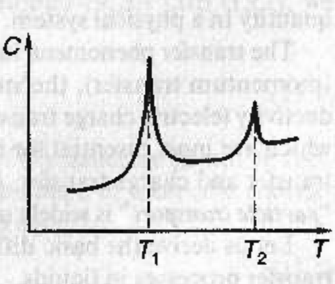


Fig. 13.6

When in membrane temperature changes, phase transitions can be seen: melting of lipids upon heating and crystallization upon cooling. Phase transitions are associated with changes in energy and that is why they can be detected, in particular, by an increase in heat capacity c with a change in temperature (fig. 13.6; at temperatures T_1 and T_2 which are phase transitions). The bilayer in liquid crystal state has lower viscosity and higher solubility of various materials than the material solid state. The thickness of the liquid crystal biolayer is less than in a solid one.

The conformation (structure) of molecules in liquid and body states is different, as it can be seen in X-ray diffraction analysis (see § 24.7). In the liquid phase, molecules of phospholipids can form strips (kinks), which can be penetrated by molecules of diffusing materials.

Displacement of a kink in this case will lead to transversal diffusion of the molecule (fig. 13.7).

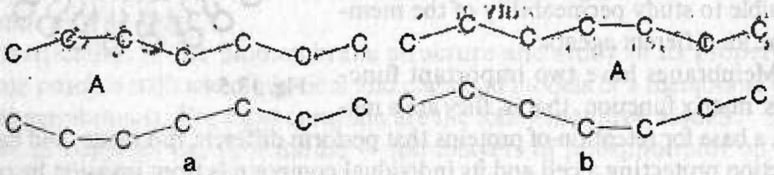


Fig. 13.7

A double phospholipid layer makes the membrane resemble a capacitor — the electric capacity of 1 mm^2 of the membrane is 5–13 nF.

§ 13.3. TRANSFER OF MOLECULES (ATOMS) ACROSS MEMBRANES

An important element of membrane functioning is their ability to pass or trap molecules (atoms) and ions. It is essential that the probability of such penetration of particles depends on the direction of their motion: into or out of a cell, and on the type of molecules and ions.

These problems are considered in the branch of physics related to ***transport phenomena***. This term covers irreversible processes causing spatial displacement (transfer) of mass, momentum, energy, charge or any other physical quantity in a physical system.

The transfer phenomena include diffusion (material mass transfer), viscosity (momentum transfer), thermal conductivity (energy transfer), electric conductivity (electric charge transfer). In this and the next paragraphs, phenomena which are most essential for biological membranes, are considered: material transfer and charge transfer. As a synonym of particle transmission, the term "***particle transport***" is widely used in biophysics.

Let us derive the basic diffusion equation (the Fick equation) considering transfer processes in liquids.

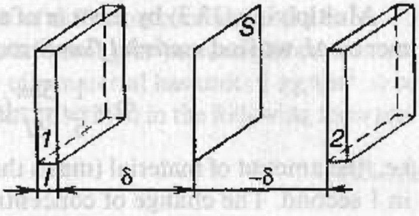


Fig. 13.8

Let molecules of a liquid move through some element with area S (fig. 13.8) in all directions. Taking into account the theory of molecular structure of a liquid (see § 9.6), it can be said that the molecules cross the element, jumping from one equilibrium position to another.

At distances equal to average displacement Δ of molecules (average distance between mole of liquid), to the right and to the left from the area element, let us plot rectangular parallelepipeds of small thickness l ($l < \delta$). The volume of each parallelepiped is equal to Sl. If n is the concentration of molecules then there are Sln molecules inside the selected parallelepipeds. Let us suppose that the concentration of molecules varies in space; in the left (*1*) selected parallelepiped, the concentration is equal to n_1, and in the right one (*2*), it is n_2. Therefore, in one parallelepiped there are Sln_1 molecules, and in the other there are Sln_2 molecules.

Due to their chaotic motion, all molecules can be conventionally represented by six groups, each of which moves along or against the direction of one of the axes of coordinates. Hence, in the direction, which is perpendicular to area element S, i.e. along the OX axis; from the first parallelepiped, $1/6\, Sln_1$ molecules jump. In direction which is opposite to the OX-axis, from the second the parallelepiped $1/6\, Sln_2$ molecules jump.

Time Δt of "flight" of these molecules through area element S can be found in the following way. Let us assume that all molecules from the selected volumes have equal average velocities $\bar{\upsilon}$. Then the molecules that have reached area element S in volume *1* or *2*, are crossing the area in the following time interval:

$$\Delta t = l/\bar{\upsilon}. \tag{13.1}$$

Substituting the expression for average velocity (9.20) into (13.1), we obtain:

$$\Delta t = \frac{l}{\delta} r, \tag{13.2}$$

where r is the average time of the "settled life" of the molecule (it can be considered as the average time of its jump). The balance of molecule transfer through area element S during time interval Δt is equal to:

$${}^1/_6 Sln_1 - {}^1/_6 Sln_2. \tag{13.3}$$

Multiplying (13.3) by mass m of a separate molecule and dividing the former by Δl, we find ***material flow*** through area element S:

$$\Phi = \frac{1}{6} \cdot \frac{Slm}{\Delta l}(n_1 - n_2), \tag{13.4}$$

i.e., the amount of material (mass) that is transferred through the area element in 1 second. The change of concentration $n_1 - n_2$ of molecules can be represented as a product of $\mathrm{d}n/\mathrm{d}x$ by distance 2δ between the selected volumes:

$$n_1 - n_2 = \frac{\mathrm{d}n}{\mathrm{d}x} \cdot 2\delta. \tag{13.5}$$

In equation (13.4), we replace Δt according to (13.2) and do $(n_1 - n_2)$ according to (13.5):

$$\Phi = \frac{1}{6}\frac{Slm\delta}{lr}\frac{\mathrm{d}n}{\mathrm{d}x} \cdot 2\delta = \frac{1}{3}S\frac{\delta^2}{r}m\frac{\mathrm{d}n}{\mathrm{d}x}. \tag{13.6}$$

The ratio of flow to the area of the area element through which it passes is called flow density:

$$J = \frac{\Phi}{S} = \frac{1}{3}\frac{\delta^2}{r}m\frac{\mathrm{d}n}{\mathrm{d}x}. \tag{13.7}$$

A product of molecule mass by its concentration is ***mass concentration*** (the ratio of molecule mass of the given component to the volume):

$$c = m \cdot n, \quad m\frac{\mathrm{d}n}{\mathrm{d}x} = \frac{\mathrm{d}c}{\mathrm{d}x}. \tag{13.8}$$

Considering this, we have from (13.7):

$$J = \frac{1}{3} \cdot \frac{\delta^2}{r} \cdot \frac{\mathrm{d}c}{\mathrm{d}x}.$$

This is the ***diffusion equation***, which is usually written as (the *Fick equation*):

$$J = -D\frac{\mathrm{d}c}{\mathrm{d}x}. \tag{13.9}$$

The sign "–" shows that the total density of material flow in diffusion tends in the direction of decreasing concentration (in the direction which is opposite to the concentration gradient), D is a ***diffusion coefficient***; as applied to the considered example of diffusion in a liquid it is equal to

$$J = \frac{1}{3}\frac{\delta^2}{r}. \tag{13.10}$$

As it can be seen from (13.10), the unit of diffusion coefficient is 1 m^2/s.

In equation (13.9), both mass concentrations (kg/m^3) and molar concentration ($mole/m^3$) can be used. Flux density of a material has units 1 $kg/(m^2 \cdot s)$ or 1 $mole/(m^2 \cdot s)$. The equation of diffusion can be written in the following form too:

$$J = -D\frac{dn}{dx}. \tag{13.11}$$

In this case, the unit of particle flux density is one part./$(m^2 \cdot s)$.

A change in the concentration of molecules in space $c = f(x)$ means that the chemical potential varies from place to place in the system [see (12.40)], so flux density of a material can be associated with the gradient of chemical potential. Let us transform formula (12.40):

$$\Delta\mu = RT\ln\frac{c_{2_i}}{c_{1_i}} = RT\ln(c_{2_i} - c_{1_i}). \tag{13.12}$$

For quite small changes in concentrations,

$$d\mu = RT\,d\ln c = RT\frac{dc}{c}.$$

or for the derivative by coordinate,

$$\frac{d\mu}{dx} = \frac{RT}{c}\frac{dc}{dx},$$

whence ***the concentration gradient*** is equal to

$$\frac{dc}{dx} = \frac{c}{RT}\frac{d\mu}{dx}. \tag{13.13}$$

Using (13.9) and (13.13), we get

$$J = -D\frac{c}{RT}\frac{d\mu}{dx}. \tag{13.14}$$

A. Einstein has shown that diffusion coefficient is proportional to temperature:

$$D = u_m RT, \quad \text{or} \quad u_m = \frac{D}{RT}. \tag{13.15}$$

Therefore, instead of (13.14), we have

$$J = -u_m c\frac{d\mu}{dx}. \tag{13.16}$$

In formula (13.15) u_m is the mobility of diffusing molecules (particles) expressed for a mole. ***The mobility of a diffusing particle*** (molecule, atom, ion, electron) u is called a coefficient of proportionality between velocity υ of the particle and force f moving the particle as other forces are acting on the latter (e.g., friction or collision with other particles) and it moves uniformly:

$$\upsilon = uf \quad \text{or} \quad u = \upsilon / f. \tag{13.17}$$

As it can be seen from (13.17), the unit of mobility is 1 m/(s · N). Values u_m and u are related by the Avogadro constant

$$u = u_m N_A. \tag{13.18}$$

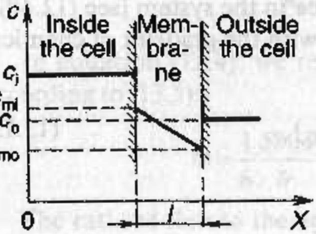


Fig. 13.9

Let us transform equation (13.9) in application to a biological membrane. We will assume that the concentration of particles diffusing through the membrane changes in it according to a linear law (fig. 13.9). The concentrations of particles inside and outside the cell are equal to c_i and c_o respectively. The concentration of these particles changes from its internal part to external one from c_{mi} to c_{mo} respectively too.

Taking into account the linear change of molecule concentration, we write down:

$$\frac{dc}{dx} = \frac{(c_{mo} - c_{mi})}{l}, \tag{13.19}$$

where l is the thickness of the membrane, then instead (13.9) we have

$$Jl = -D(c_{mo} - c_{mi}) = D(c_{mi} - c_{mo}). \tag{13.20}$$

It is more accessible practically to determine the concentration of particles not inside the membrane (c_{mi} and c_{mo}), but outside it: in cell (c_i) and outside cell (c_o). In the membrane, the ratio of boundary values of concentrations is not considered equal to the ratio of concentrations in layers adjacent to the membrane: $c_{mo}/c_{mi} = c_o/c_i$, whence:

$$\frac{c_{mo}}{c_o} = \frac{c_{mi}}{c_o} = k, \tag{13.21}$$

where k is the coefficient of material (particles) distribution between the membrane and the environment (usually, at the initial stage). From (13.21), it follows

$$c_{mo} = kc_o \quad \text{and} \quad c_{mi} = kc_i. \tag{13.22}$$

Substituting (13.22) into (13.20), we have

$$J = \frac{Dk}{l}(c_i - c_o).$$

Let

$$P = Dk/l, \tag{13.23}$$

where P is the permeability coefficient. As a result, we obtain an equation for the flow density of a material during diffusion through a biological membrane:

$$J = P(c_i - c_o). \tag{13.24}$$

§ 13.4. NERNST–PLANCK EQUATION. TRANSFER OF IONS ACROSS MEMBRANES

As it will be shown in § 13.6, there is potential difference on the membrane, hence, there is an electric field in the membrane too. It affrects the diffusion of charged particles (ions and electrons). There is a known relationship between field intensity and the potential gradient (see § 14.1):

$$E = -\frac{d\varphi}{dx}. \tag{13.25}$$

If the valence of an ion is Z, then its charge is equal to $Z \cdot e$. One ion $f = Ze\frac{d\varphi}{dx}$ is influenced by force. The force acting on 1 mole of ions is:

$$fN_A = -ZeN_A\frac{d\varphi}{dx} = -ZF\frac{d\varphi}{dx}, \tag{13.26}$$

where F is the Faraday constant, $F = eN_A$.

The velocity of the directional motion of ions is proportional to the acting force:

$$\upsilon = u_m fN_A = -u_m ZF\frac{d\varphi}{dx}. \tag{13.27}$$

To find the flow of ions, let us select the volume of the electrolyte (fig. 13.10) in the form of a rectangular parallelepiped with an edge, which is numerically equal to the actual velocity of ions. In the parallelepiped, all ions will pass in 1 s through area element S. It will be exactly flow Φ. The number of these ions can be found multiplying the parallelepiped volume (υS) by the ion molar concentration c:

$$\Phi = \upsilon Sc. \tag{13.28}$$

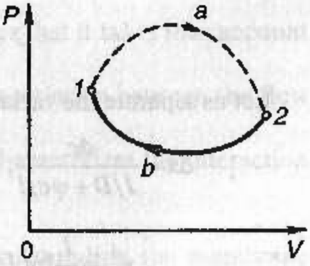


Fig. 13.10

We will find the flow density using formulas (13.27) and (13.28):

$$J = \frac{\Phi}{S} = \upsilon c = -u_m ZFc\frac{d\varphi}{dx}. \tag{13.29}$$

In general, ion transfer is determined by two factors: non-uniformity of their distribution, i.e. the concentration gradient [see (13.9)], and the effect of the electric field [see (13.29)]:

$$J = -D\frac{dc}{dx} - u_m ZFc\frac{d\varphi}{dx}. \tag{13.30}$$

It is the ***Nernst–Planck equation***. Using the expression for mobility (13.15), we transform the equation (13.30):

$$J = -D\frac{dc}{dx} - \frac{D}{RT}ZFc\frac{d\varphi}{dx} = -D\left(\frac{dc}{dx} + \frac{ZFc}{RT}\frac{d\varphi}{dx}\right). \tag{13.31}$$

It is another form of the Nernst–Planck equation.

Let us use the Nernst–Planck equation to determine the dependence of the diffusion flow density on ion concentration and electric field intensity. Let us assume that the state is stationary (flow density F is constant). The electric field in the membrane is assumed to be homogeneous, hence, the electric field intensity is constant and the potential varies with distance linearly. It makes it possible to consider that $\frac{d\varphi}{dx} = \frac{\varphi_m}{l}$, where φ_m is the potential difference on the membrane.

Let us simplify the summand in equation (13.31):

$$\frac{ZFc}{RT}\frac{d\varphi}{dx} = \frac{ZFc}{RT}\frac{\varphi_m}{l} = \frac{\psi}{l}c,$$

where

$$\psi = \frac{ZFc}{RT}\varphi_m \tag{13.32}$$

is an auxiliary quantity (dimensionless potential). Taking into account (13.32) we obtain Nernst–Planck equation in the form

$$J = -D\left(\frac{dc}{dx} + \frac{\psi}{l}c\right). \tag{13.33}$$

Let us separate the variables and integrate the equation:

$$dx = \frac{dc}{J/D + \psi c/l}; \quad \int_0^l dx = -\int_{c_{m2}}^{c_{m0}} \frac{dc}{J/D + \psi c/l} = \int_{c_{m0}}^{c_{m2}} \frac{dc}{J/D + \psi c/l},$$

$$x\Big|_0^l = \frac{l}{\psi}\ln(J/D + \psi c/l)\Big|_{c_{mo}}^{c_{mi}} = \frac{l}{\psi}\ln\frac{J/D + \psi c_{mi}/l}{J/D + \psi c_{mo}/l}.$$

From here, we find

$$\psi = \ln \frac{J/D + \psi c_{mi}/l}{J/D + \psi c_{mo}/l} \tag{13.34}$$

Finding the antilogarithm of (13.34), we obtain:

$$e^{\psi} = \ln \frac{J/D + \psi c_{mi}/l}{J/D + \psi c_{mo}/l},$$

whence

$$e^{\psi}\frac{J}{D} + e^{\psi}\frac{\psi}{l}c_{mo} = \frac{J}{D} + \frac{\psi}{l}c_{mi}, \quad \frac{J}{D}(e^{\psi} - 1) = \frac{\psi}{l}(c_{mi} - e^{\psi}c_{mo}),$$

$$J = \frac{D\psi}{l}\frac{c_{mi} - e^{\psi}c_{mo}}{e^{\psi} - 1}. \tag{13.35}$$

Let us transform formula (13.35), taking into account expressions (13.22) and (13.23):

$$J = \frac{D\psi}{l}\frac{kc_i - e^{\psi}kc_o}{e^{\psi} - 1} = \frac{D\psi k}{l}\frac{c_i - e^{\psi}c_o}{e^{\psi} - 1} = D\psi\frac{c_i - e^{\psi}c_o}{e^{\psi} - 1}. \tag{13.36}$$

Generally speaking, formula (13.36) is valid both for positive ($Z > 0$, $\psi > 0$), and for negative ($Z < 0$, $\psi < 0$) ions. Nevertheless, for negative ions, it is expedient to modify this expression substituting a negative value of dimensionless potential into it:

$$J_{(-)} = -P\psi\frac{c_i - e^{-\psi}c_o}{e^{-\psi} - 1}.$$

Let us divide the numerator and denominator of this expression by $e^{-\psi}$:

$$J_{(-)} = -P\psi\frac{c_o - e^{\psi}c_i}{e^{\psi} - 1}. \tag{13.37}$$

Using this formula, it is necessary to remember that it takes into account negative values of Z and ψ.

Equations (13.36) and (13.37) determine the association between the flow density of a stationary ion flow with three values:

1) permeability of membranes for this ion; it characterizes the interaction of membrane structures with the ion;
2) electric field;
3) ion concentrations in the water solution surrounding the membrane (c_i and c_o).

Let us analyze particular cases of equation (13.36):

a) $\psi = 0$, which means either $Z = 0$ (neutral particles), absence of electric field in the membrane ($\varphi_m = 0$), or both cases:

$$\lim_{\psi \to 0} J = P \lim \frac{\psi}{e^{\psi} - 1} \lim_{\psi \to 0}(c_i - e^{\psi} c_o).$$

Let us find the limits for individual factors:

1. $\lim\limits_{\psi \to 0} \frac{\psi}{e^{-\psi} - 1} = \frac{0}{0}$. This indeterminate form can be evaluated by the l'Hôspital rule:

$$\lim_{\psi \to 0} \frac{\psi}{e^{-\psi} - 1} = \lim_{\psi \to 0} \frac{1}{e^{-\psi}} = 1.$$

2. $\lim\limits_{\psi \to 0}(c_i - e^{\psi} c_o) = c_i - c_o.$

Whence, we obtain equation (13.24) just as we expected:

$$J = P(c_i - c_o);$$

b) the same concentration of ions on different sides of the membrane ($c_i = c_o = c$) in the presence of an electric field:

$$J = -P\psi c.$$

This corresponds to the case of electrical conductivity in an electrolyte (see § 15.3). For neutral particles ($Z = 0$ and $\psi = 0$) $J = 0$;

b) if the membrane is impenetrable for particles ($P = 0$) then, naturally the flux density is zero.

§ 13.5. ACTIVE TRANSPORT

Transfer phenomena (see § 13.3 and 13.4) belong to ***passive transport***, which is diffusion of molecules and ions in the direction of their lower concentration, motion of ions in accordance with the direction of the force acting on them on the part of electric field.

Passive transfer is not associated with expenditure of chemical energy, it occurs due to particles moving towards a lower electrochemical potential (see § 12.5). Along with passive transport, in cell membranes there is transfer of molecules and ions in the direction of a greater electrochemical potential (molecules are transferred to an area of their higher concentration; ions are transferred against the forces acting on them on the part of the electric field).

Transfer is called ***active transport***, when it occurs at the expense of energy and is not diffusion. Membrane systems contributing to formation of ion gradients K^+ and Na^+, are called sodium-potassium pumps or simply sodium pumps.

Sodium-potassium pumps are parts of cytoplasm membranes, they function at the expense of the hydrolysis energy of ATP molecules with the formation of ADP molecules and inorganic phosphate (Φ_i):

$$\text{ATP} = \text{ADP} + \Phi_i.$$

A sodium-potassium pump works in a reversible way: ion concentration gradients contribute to synthesis of ATP from ADP and Φ_i molecules:

$$\text{ADP} + \Phi_i = \text{ATP}.$$

The work that is done for transferring one mole of a material from the area with a lower concentration c_{1i} of this component into the area with a higher concentration c_{2i}, can be found as a change in Gibbs energy or a change in the chemical potential [see (12.39)]:

$$A = \Delta\mu = RT \ln \frac{c_{2i}}{c_{1i}}. \tag{13.38}$$

Thus, if inside the cell, the concentration of K^+ is taken to be 50 times higher than that outside the cell then according to (13.38), at a temperature of 36 °C, we obtain

$$A = 8.31\ \text{J/(mole}\cdot\text{K)}\cdot 309\text{K} \ln 50 = 10\ \text{kJ/mole}.$$

The mechanism of the sodium-potassium pump action is not clear enough, but it is essential that it functions under the condition of conjugating potassium and sodium. It means absence of active transfer of Na^+ ions from the cell if there are no K^+ ions in the environment and no transfer of K^+ ions into the cell, if there are Na^+ ions in it. In other words sodium ions activate the sodium-potassium pump on the inner surface of cell membrane, and potassium ions do the same on the outside surface.

Sodium-potassium pump transfers three sodium ions from the cell to the environment in exchange for t transfer of two potassium ions into the cell. This creates and maintains a difference of potentials on the membrane; the inner part of the cell has a negative charge (see § 13.7).

§ 13.6. VARIETIES OF PASSIVE TRANSFER OF MOLECULES AND IONS ACROSS BIOLOGICAL MEMBRANES

Simple diffusion through the lipid layer is shown schematically in fig. 13.11, It meets the Fick equation for molecules (13.9) or, in more common case, for neutral and charged particles, meets the Nernst–Planck equation (13.31). In a living cell, such diffusion ensures passage of oxygen and carbon dioxide.

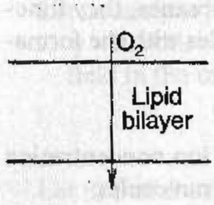


Fig. 13.11

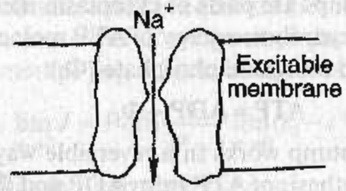


Fig. 13.12

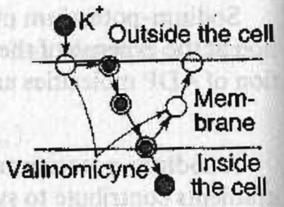


Fig. 13.13

A number of medications and poisons also pass through the lipid layer according to the diagram shown in the figure. As it is noted in § 13.2, certain conformation of phospholipids facilitates diffusion across a membrane due to motion of kinks.

However, such simple diffusion runs quite slowly and cannot supply the cell with the required amount of nutrients. Therefore, there are other mechanisms of passive transfer of material in a membrane, these include diffusion through the canal (stoma) (fig. 13.12) and facilitated diffusion (in combination with a transporter).

A ***stoma***, or ***canal***, is a part of a membrane including protein molecules and lipids; a stoma forms a passage. The passage lets in through the membrane not only small molecules, like those of water, but also greater ions. Diffusion through the stomas is also described by diffusion equations; however, presence of stomas increases permeability *P*. Canals can be selective with respect to different ions; it will be manifested in difference of permeability to different ions.

In addition, transfer of ions by special transferring molecules (fig. 13.13) eases diffusion. The most known ability of a molecule of valinomycin (an antibiotic) to transfer a potassium ion across model bilayer membranes. This molecule captures a potassium ion, forms a lipid-soluble complex, and passes across the membrane. Thanks to their ability to transfer ions across membranes, valinomycin and other related compounds are called ***ionophores***.

With the help of carriers, transport can be carried out as relay-race ion motion. In this case, transport molecules form a temporary chain across the membrane and pass the diffusing molecule one to another.

Diffusion of a valinomycin molecule complex and potassium ion is also described by the general diffusion equation.

Ion transfer across membranes was studied by V.F. Antonov.

§ 13.7. RESTING POTENTIAL

The cell surface of membrane is not equally permeable to different ions. In addition, the concentration of any particular ions varies on different sides of the membrane inside the cell, the most favorable composition of ions is maintained. In a normally functioning cell, these factors generate a potential difference (***resting potential***) between the cytoplasm and environment.

The main contribution to creation and maintenance of the rest potential is made by ions of Na^+, K^+ and Cl^-. The total flow density of these ions, taking into account their signs, is

$$J = J_{Na^+} + J_{K^+} - J_{Cl^-}. \tag{13.39}$$

In the stationary state, the total flow density is zero, i.e. the number of different ions passing across the membrane into the cell per time unit is equal to the number of ions leaving the cell across the membrane: $J = 0$.

For a flow density of positive ions (Na^+ and K^+) we write down expression (13.36), and for the flow density of negative ions of chlorine we do expression (13.37). Summing up these flows, we obtain:

$$\begin{gathered} P_{Na}\Psi\frac{[Na^+]_i - e^{\Psi}[Na^+]_0}{e^{-\Psi}-1} + P_{K}\Psi\frac{[K^+]_i - e^{\Psi}[K^+]_0}{e^{-\Psi}-1} + \\ + P_{Cl}\Psi\frac{[Cl^-]_0 - e^{\Psi}[Cl^-]_i}{e^{-\Psi}-1} = 0. \end{gathered} \tag{13.40}$$

In this, in square brackets $[\ ]_i$ and $[\ ]_0$, concentrations of ions inside and outside the cell respectively are indicated. Cancelling (13.40) by $\frac{\Psi}{e^{\Psi}-1}$, expanding the expressions and regrouping them, we obtain:

$$\begin{gathered} P_{Na}[Na^+]_i + P_K[K^+]_i + P_{Cl}[Cl^-]_0 = \\ = e^{\Psi}\{P_{Na}[Na^+]_0 + P_K[K^+]_0 + P_{Cl}[Cl^-]_i\}, \end{gathered}$$

or

$$e^{\Psi} = \frac{P_{Na}[Na^+]_i + P_K[K^+]_i + P_{Cl}[Cl^-]_0}{P_{Na}[Na^+]_0 + P_K[K^+]_0 + P_{Cl}[Cl^-]_i}.$$

Finding the logarithm of this expression, we find:

$$\psi = \ln\frac{P_{Na}[Na^+]_i + P_K[K^+]_i + P_{Cl}[Cl^-]_0}{P_{Na}[Na^+]_0 + P_K[K^+]_0 + P_{Cl}[Cl^-]_i}. \tag{13.41}$$

If we return from the dimensionless potential to the electric one [see (13.32)], then from (13.41) we have:

$$\varphi_m = \frac{RT}{F}\ln\frac{P_{Na}[Na^+]_i + P_K[K^+]_i + P_{Cl}[Cl^-]_0}{P_{Na}[Na^+]_0 + P_K[K^+]_0 + P_{Cl}[Cl^-]_i}. \tag{13.42}$$

This is ***the Goldman–Hodgkin–Katz equation***.

Ion pumps, which are active transport systems, generated different concentrations of ions inside and outside a cell. We can say that a resting potential exists thanks to active transport.

As an example of using the Goldman–Hodgkin–Katz equation, we calculate the resting potential of a giant squid axon.

The concentration values are given in Table 13.1.

Table 13.1

Ion	Concentration, mmol per kg of H_2O	
	inside the cell	outside the cell
K^+	340	10.4
Na^+	49	463
Cl^-	114	592

Ion permeability depends on the state of a body very much. At rest in physiological conditions, the ratio of permeability coefficient is equal to

$$P_K : P_{Na} : P_{Cl} = 1 : 0.04 : 0.45.$$

Therefore, the main contribution to the resting potential is made only by ions of K^+ and of Cl^-. According to (13.42), at 30 °C we have

$$\varphi_m = \frac{8.3 \cdot 303}{9.6 \cdot 10^4} \ln \frac{304 + 0.45 \cdot 592}{10.4 + 0.45 \cdot 114} \approx 59.7 \text{ mV}.$$

This agrees well with the experimental value of 60 mV. From the Goldman–Hodgkin–Katz equation, one can deduce the Nernst equation (12.46) for the equilibrium state. At all that, we should neglect permeability of all ions (e.g., $P_{Na} = 0$ and $P_{Cl} = 0$) except for ions of the same kind ($P_K \neq 0$). Then for potassium ions

$$\varphi_m = -\frac{RT}{F} \ln \frac{[K^+]_i}{[K^+]_0}.$$

§ 13.8. ACTION POTENTIAL AND ITS PROPAGATION

Upon excitement, the difference of potentials between a cell and the environment changes, and an ***action potential*** is generated (fig. 13.14). It resembles aperiodic processes that take place when a capacitor is charged and discharged (see § 18.1). In nerve fibers, an action potential propagates.

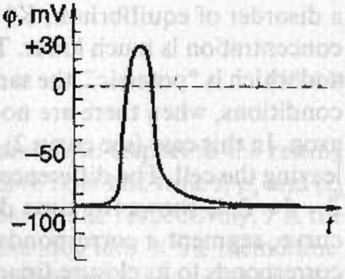


Fig. 13.14

During occurrence of an action potential, measurement of membrane resistance showed that it changes, repeating the time dependence of the action potential according to the formula. To answer the question for which ions membrane permeability changes, it should be noted that the action potential generates a short-term increase in the potential inside the cell (fig. 13.14). The negative potential relative to the environment becomes positive. If we calculate the equilibrium potentials on the membrane of the squid axon using the Nernst equation (12.46), we obtain for the values of −90, +46, and −29 mV for K^+, Na^+ and Cl^- ions respectively. Since when membrane permeability changes for any ion, the ion will pass through the membrane tending to create an equilibrium state, the numerical data show that Na^+ ions pass into the cell generating a positive potential there. Therefore, when the cell is excited, in the initial period the permeability of membranes increases just for sodium ions.

It is possible to measure membrane permeability for any ion (in other words, the electrical conductivity or resistance of the membrane for this ion) if using Ohm's law we find the ratio of current intensity to voltage, or vice versa. Practical implementation of this problem is complicated by the fact that during excitation, membrane permeability (electrical resistance) changes in the course of time. It generates redistribution of voltage in the circuit, and the difference of potentials on the membrane changes. Hodgkin, Huxley and Katz were able to create an experiment fixing a definite value of potential difference on a membrane, which permitted them to measure ion currents and, therefore, measure membrane ion permeability (resistance).

The results of this interesting work are shown in fig. 13.15. Curve *1* corresponds to time dependence of the intensity of ion current obtained by changing the potential on the membrane to +56 mV (the resting potential is −60 mV) and passing through the membrane of a giant squid axon. The direction of the current is negative; this means passage of positive ions across the membrane into the cell. At the initial period of excitation, as already determined, this current is generated by passage of Na^+ ions into the cell, where their concentration is much less than outside of it. Naturally, with such

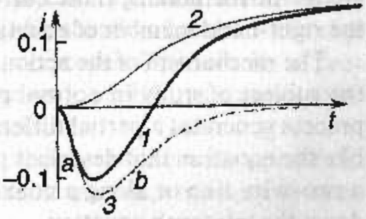


Fig. 13.15

a disorder of equilibrium, K^+ ions will start to move outwards, where their concentration is much lower. To find out which part of the current is "sodic" and which is "potassic", the same excitation can be carried out, but in artificial conditions, when there are no Na^+ ions in the environment surrounding the axon. In this case (see curve 2), the current is generated only thanks to K^+ ions leaving the cell. The difference between two curves is shown by in third curve *3 = 1 – 2*; it shows the time dependence of the ion current of Na^+. In this curve, segment *a* corresponds to opening of sodium canals, and segment *b* corresponds to its closure (inactivation).

Ion channels are of protein origin (see fig. 13.3 and 13.11). They pass ions of different kind selectively. A channel can be closed (blocked) by molecules of poisons, its permeability depends on action of some medications. That is why the theory of ion channels in membranes is an important part of molecular pharmacology.

Having analyzed experimental data by methods of mathematical simulation, Hodgkin and Huxley could obtain curves that describe action potentials of various excitable structures. In particular, the scientists derived the expression for density of the total current passing across a membrane:

$$j = C\frac{d\varphi}{dt} + (\varphi - \varphi_K)\bar{g}_K n^4 + (\varphi - \varphi_{Na})\bar{g}_{Na} m^3 h + (\varphi - \varphi_y)\bar{g}_y, \qquad (13.43)$$

where C is electrical capacity of the membrane per unit of its area; φ is action potential (difference of potentials between external and internal membrane surfaces); φ_K, φ_{Na}, φ_y are equilibrium potentials for K^+, Na^+ and all other ions respectively; $\bar{g}_K$, $\bar{g}_{Na}$, $\bar{g}_y$ are specific conductivities of the membrane for the corresponding ions at completely open channels; n is the part of active (open) K^+channels; m is the same for Na^+; h is the part of sodium channels, which cannot be inactivated (closed). Thus, the density of a current passing across a membrane is determined by capacitive current (displacement current) caused by a change in the charge on the outer and inner surfaces of the membrane $\left(C\frac{d\varphi}{dt}\right)$ and conductivity currents caused by passage of different ions across the membrane; these currents are expressed by the other three terms in the right-hand member of equation (13.43).

The mechanism of the action potential propagation along the nerve fiber is the subject of study in normal physiology. Mathematical description of this process generates a partial differential equation (telegraph equation), which is like the equation that describes propagation of an electromagnetic wave along a two-wire line or along a coaxial cable. There are several forms of writing down the telegraph equation.

Here is one of them:

$$\frac{\partial^2 \varphi}{\partial x^2} = \frac{2\rho_a}{r}\left(C\frac{\partial \varphi}{\partial t} + \frac{\varphi}{\rho_m l}\right). \tag{13.44}$$

In this, φ is an intracellular potential measured with respect to the resting potential. It depends on distance x along the nerve fiber and time t; ρ_a and ρ_m are specific resistances of the axoplasm and membrane respectively; r is the nerve fiber radius; C_m is electrical capacity of the area unit of the membrane.

Equation (13.44), as a partial differential equation, is not considered in this course.

Along with some formal analogy between the excitation wave (propagating an action potential along a nerve fiber) and an electromagnetic wave in a two-wire line, there is essential fundamental difference between these waves.

An electromagnetic wave, propagating in a medium, attenuates because it loses its energy. There is ***dissipation of oscillation energy***, i.e., conversion of the energy of oscillations (waves) into energy of molecular thermal motion.

The energy source of an electromagnetic wave is its source itself: generator, spark, etc.

An excitation wave does not attenuate, as it receives energy from the medium in which it spreads (in this example, it is the energy of a charged membrane).

Waves that receive energy from the medium for their propagation are called ***autowaves***, and the medium is called ***active***.

Thus, an action potential spreads along the nerve fiber in the form of an autowave. The active medium is excitable cells.

Calculations show that along smooth unmyelinated nerve fibers, the velocity of excitation propagation is approximately proportional to the square root of their radius ($\upsilon \approx \sqrt{r}$). In some invertebrates, quite high velocities of propagation of action potential (20–30 m per second) are due to a great diameter of their fibers (up to 1 mm).

In vertebrates, the velocity of excitation propagation is generated by myelination of fibers. Specific resistance of myelin is greater than in other biological membranes; the thickness of the myelin sheath is hundreds of times greater than the thickness of an ordinary cell membrane. The velocity of propagation is proportional to both the thickness and specific resistance of the membrane, so the velocity is quite high in vertebrates.

A lesion of the myelin sheath generates a disorder of the action potential propagation along the nerve fiber and, as a consequence, to severe nervous diseases.

Section 4

ELECTRODYNAMICS

Electric and magnetic phenomena are associated with a special form of manifestation of matter — electric and magnetic fields and their interactio These fields are generally so interdependent that it is customary to refer to the electromagnetic field.

Electromagnetic phenomena have three directions of medical and biological applications. The first of them is understanding the electrical processe taking place in the body, as well as studying electrical and magnetic properti of biological media, for instance, the physics underlying electrocardiograph magnetobiology, rheography, electric conductivity of biological tissues and fluids, etc. The second direction is related to understanding the mechanism of electromagnetic fields impact on the body. This impact can serve as therapeutic, industrial, or climatic factor. The third direction is instrumental. Electrodynamics is the theoretical basis of electronics and, in particular, medica electronics. In this respect, electrodynamics is more important for medicin because many non-electrical parameters of biological systems, such as temperature, are measured by conversion into an electric signal to facilitate measurement and recording.

Chapter 14

Electric field

An electric field is a kind of matter by means of which a force action is carried out on electric charges located in this field. The characteristics of the electric field generated by biological structures are a source of information about the state of the body.

§ 14.1. INTENSITY AND POTENTIAL AS CHARACTERISTICS OF ELECTRIC FIELD

The force characteristic of an electric field is its intensity equal to the ratio of the force acting at a given point of the field on a point charge, to the value of the charge:

$$\vec{E} = \vec{F}/q. \tag{14.1}$$

Intensity is a vector whose direction coincides with the direction of the force acting at this point of the field on the positive point charge.

The electric field intensity is analytically determined at arbitrary points by the following three equations:

$$E_x = f_1(x,y,z);\ E_y = f_2(x,y,z);\ E_z = f_3(x,y,z), \tag{14.2}$$

where E_x, E_y and E_z are vector projections onto the corresponding coordinate axes introduced to describe the field. Electric field can be conveniently represented by lines of force, tangents to which coincide with the direction of the vector of intensity in the corresponding points of the field.

Usually these lines are drawn out with such a density that the number of lines passing through a unit area element perpendicularly to them is equal to the value of the electric field intensity at the location of the area element.

A *potential* is an energy characteristic of an electric field.

Let us imagine that an electric charge q is moving in an electric field along trajectory *1–a–2* (fig. 14.1) while the forces of the field perform work that can be expressed through intensity [see (14.1)]:

$$A = \int_1^2 E_l q \cdot \mathrm{d}l = q\int_1^2 E_l \cdot \mathrm{d}l, \tag{14.3}$$

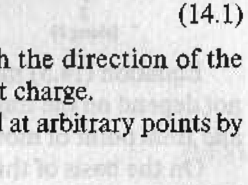


Fig. 14.1

where dl is elementary displacement; E_l is projection of E onto the direction of dl. **Let us prove that the work of electrostatic field forces does not depend on the trajectory along which the charge is moving in this field.** ***Fields with such a property are called potential fields.***

Let charge q move along closed trajectory *1–a–2–b–1* (fig. 14.1). Since the field is electrostatic, the position of the charges creating the field has not changed, and the potential energy depending of their mutual position is the same. Therefore, **the work of the electrostatic field forces moving a charge along a closed trajectory is zero**[1]:

$$A = q\oint E_l \mathrm{d}l = q\int_1^2 E_l \mathrm{d}l + q\int_2^1 E_l \mathrm{d}l = 0. \qquad (14.4$$

Since the forces acting on charge q are determined by its position in the field, when a charge moves along the same trajectory in opposite directions expressions for the work of the field forces differ only by the sign:

$$q\underset{(\text{along } b)}{\int_2^1} E_l \mathrm{d}l = -q\underset{(\text{along } b)}{\int_1^2} E_l \mathrm{d}l; \quad q\underset{(\text{along } a)}{\int_2^1} E_l \mathrm{d}l = -q\underset{(\text{along } b)}{\int_1^2} E_l \mathrm{d}l. \qquad (14.5$$

Equation (14.5) means that the work of forces of an electrostatic field does not depend on the trajectory of a charge, but depends on the charge, the initial and final point of movement, and field intensity.

On the basis of this property, a concept of difference of potentials, or voltage, is introduced.

The difference of potentials between two points of the field is the ratio of the work done by the field forces moving a positive point charge from one point of the field to another, to this charge:

$$U_{12} = \varphi_2 - \varphi_1 = \frac{A}{q} = \frac{q\int_1^2 E_l \mathrm{d}l}{q} = \int_1^2 E_l \mathrm{d}l, \qquad (14.6)$$

where φ_1 and φ_2 are potentials at points *1* and *2* of the electric field. The difference of potentials between two points depends on the actual position of the selected points and the electric field intensity [see (14.6)].

Along with a difference of potentials, the concept of potential is used as a characteristic of the electric field. However, for a given field point, it is unambiguous only if the potential of any arbitrary field point is given. In practice, it is accepted that the potential of conductors connected to earth ground, or the

---

[1] Radiative losses are assumed to be negligible.

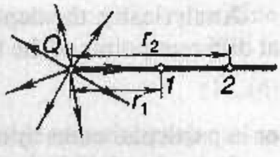


Fig. 14.2

potential of a chassis on which a radio set is mounted (in both cases, we are talking about ground), are equal to zero. In theoretical problems, the potential of infinitely remote points is usually considered to be equal to zero.

Let us calculate the potential of the field of a point charge in a homogenous isotropic dielectric with relative dielectric permeability ε_r (fig. 14.2). Let points *1* and *2* be located in the same line of force at distances r_1 and r_2 from the field source, which is charge Q, respectively. Integrating expression (14.6) along the segment *1–2*, taking into account that $E_l = E = Q/(4\pi\varepsilon_r\varepsilon_0 r^2)$ (for a point charge) and $\mathrm{d}r = \mathrm{d}l$:

$$\varphi_2 - \varphi_1 = \int_{r_1}^{r_2} E\,\mathrm{d}r = \frac{Q}{4\pi\varepsilon_r\varepsilon_0}\int_{r_1}^{r_2}\frac{\mathrm{d}r}{r^2} = \frac{Q}{4\pi\varepsilon_r\varepsilon_0 r_1} - \frac{Q}{4\pi\varepsilon_r\varepsilon_0 r_2}, \qquad (14.7)$$

where $\varepsilon_0 \approx 8.84 \cdot 10^{-12}$ F/m is the electric constant.

Let us assume that the potential at an infinitely remote point is zero: $\varphi_2 \to 0$ at $r_2 \to \infty$. Then from (14.7) we obtain:

$$\varphi = \frac{Q}{4\pi\varepsilon_r\varepsilon_0 r_1},$$

or in a more general form:

$$\varphi = \frac{Q}{4\pi\varepsilon_r\varepsilon_0 r}. \qquad (14.8)$$

There could be other suggestions about the nature of potential at the infinitely remote point, but the above assumption has led us to the simplest expression (14.8), by which the potential of a field of point charge is usually calculated.

Potentials at various points can be visually represented in the form of ***surfaces of equal potential*** (***equipotential surfaces***). Usually, equipotential surfaces are drawn so that they differ from adjacent surfaces by the equal value of potential. Fig. 14.3 shows equipotential surfaces[1] (dash lines) and lines of forces (continuous lines) of a field of two equal point charges of different signs.

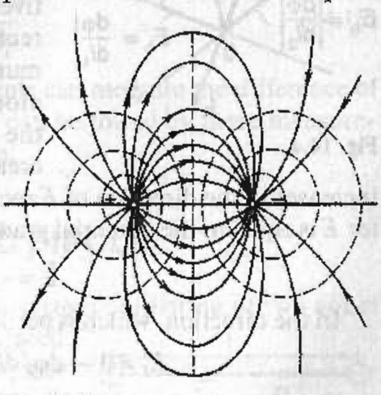


Fig. 14.3

---

[1] In the drawing plane, equipotential surfaces are represented by equipotential lines.

Analytically, the dependence of the electric potential on the coordinate at different points of the field is given by the equation:

$$\varphi = f(x,y,z) \qquad (14.9$$

or in particular cases by other dependencies like (14.8).

Since the electric field intensity is determined through force, and the potential is determined through the work of field forces, these characteristics are connected like force and work. The integral dependence is already given b formula (14.6), or

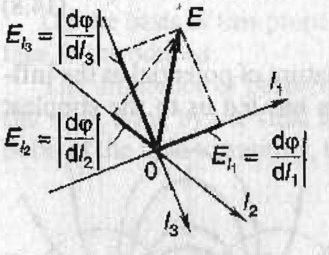
$$U_{21} = \varphi_2 - \varphi_1 = -\int_1^2 E_l \mathrm{d}l. \qquad (14.1$$

In this, the upper limit of the integral, as it follows from mathemati is matched in the left-hand member by minuend φ_2, the lower one is subtrahend φ_1.

Let us obtain the differential connection between E and φ. Let us assum that points *2* and *1* are located arbitrarily close, and then from (14.10), we obtain

$$\mathrm{d}\varphi = -E_l\,\mathrm{d}l, \text{ or } E_l = -\mathrm{d}\varphi/\mathrm{d}l. \qquad (14.11)$$

The directional derivative of the potential $\mathrm{d}\varphi/\mathrm{d}l$ characterizes the ratio of the potential increment to the corresponding length in direction l; E_l is the projection of vector E onto this direction.



$E_{l_3} = \left|\frac{\mathrm{d}\varphi}{\mathrm{d}l_3}\right|$
E
l_1
$E_{l_2} = \left|\frac{\mathrm{d}\varphi}{\mathrm{d}l_2}\right|$
0
$E_{l_1} = \left|\frac{\mathrm{d}\varphi}{\mathrm{d}l_1}\right|$
l_2
l_3

Fig. 14.4

The meaning of formula (14.11) is clear fron fig. 14.4. At point O, the vector E, is drawn; it is projected onto directions l_1, l_2 and l_3. These projections are numerically equal to the derivatives of the potential in the corresponding directions of $\mathrm{d}\varphi/\mathrm{d}l_1$, $\mathrm{d}\varphi/\mathrm{d}l_2$, $\mathrm{d}\varphi/\mathrm{d}l_3$. The maximum potential change per unit of length occu along the straight line coinciding with the minus sign means that the potential decreases in the direction of E most quickly and increases in the direction of E most quickly respectively. We can say that vector E is equal to the potential gradient taken with the opposite sign:

$$\vec{E} = -\mathrm{grad}\,\varphi. \qquad (14.12)$$

In the direction, which is perpendicular to the line of force, we have

$$E_l = 0 \rightarrow \mathrm{d}\varphi/\mathrm{d}l = 0 \rightarrow \varphi = \mathrm{const}. \qquad (14.13)$$

From this, it follows that **lines of force and equipotential surfaces are mutually perpendicular**. If a field is uniform, like a field inside a flat capacitor, then, from formula (14.6), we find that for two points located on the same line of force at distance l,

$$E=(\varphi_2-\varphi_1)/l=U_{12}/l. \tag{14.14}$$

Taking into account (14.11) and (14.9) it is possible to write projections of the electric field intensity vector onto three coordinate axes: homogeneous

$$E_x=-\mathrm{d}\varphi/\mathrm{d}x,\ E_y=-\mathrm{d}\varphi/\mathrm{d}y,\ E_z=-\mathrm{d}\varphi/\mathrm{d}z. \tag{14.15}$$

Then the intensity is determined by the formula:

$$E=\sqrt{E_x^2+E_y^2+E_z^2}, \tag{14.16}$$

and direction E is determined by values of cosines of angles between this vector and the coordinate axes:

$$\cos(\vec{E},x)=E_x/E;\ \cos(\vec{E},y)=E_y/E;\ \cos(\vec{E},z)=E_z/E.$$

If the field is generated by N point charges, the intensity at a point can be calculated as the vector sum of the field intensities generated by each charge at that point separately (superposition principle):

$$\vec{E}=\sum_{i=1}^{N}\vec{E}_i, \tag{14.17}$$

and the electric potential at this point is the algebraic sum of potentials from each charge, assuming that the potential of infinitely remote points is equal to zero:

$$\varphi=\sum_{i=1}^{N}\varphi_i=\sum_{i=1}^{N}\frac{Q}{4\pi\varepsilon_r\varepsilon_0 r_i}. \tag{14.18}$$

Available electricity measuring instruments can measure the difference of potentials, rather than intensity. Intensity can be found by these measurements using the connection between E and φ.

§ 14.2. ELECTRIC DIPOLE

An electric dipole (simply a dipole) is a system consisting of two equal point electric charges with opposite signs located at some distance from each other (a dipole shoulder). The main characteristic of a dipole (fig. 14.5) is its electric, or dipole moment — a vector equal

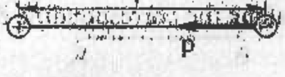


Fig. 14.5

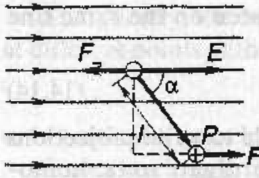


Fig. 14.6

to the product of the charge by the dipole shoulde directed from the negative charge to the positiv one:

$$\vec{p} = q\vec{l}. \qquad (14.19)$$

A unit of electric moment of dipole is a *coulomb-meter.*

Let us place a dipole in a uniform electric field with intensity $\vec{E}$ (fig. 14.6).

Each dipole charge is exposed to force $\vec{E}_+ = q\vec{E}$ and $\vec{E}_- = -q\vec{E}$; these force are directed oppositely and generate a moment of a pair of forces. As it can be seen from the figure, it is equal to

$$M = qE_l \sin\alpha = \mathrm{p}E\sin\alpha, \qquad (14.20)$$

or in the vector form

$$\vec{M} = \vec{p}\times\vec{E}. \qquad (14.21)$$

Thus, a torque depending on the electric moment, orientation of the dipol in the field and field intensity act on the dipole in a homogenous electric field.

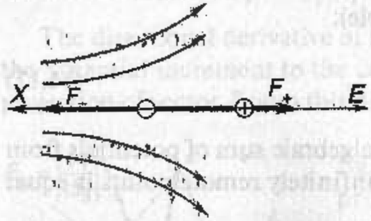


Fig. 14.7

Let us consider now a dipole in a non-uniform electric field. Suppose that the dipole is located along a line of force whos direction is opposite to the X-axis (fig. 14.7). It is under the action of forces

$$\vec{E}_+ = q\vec{E}_- \text{ and } \vec{E}_- = -q\vec{E}_-,$$

where E_+ and E_- are field intensities, respectively, at the location of positive and negative charges (in fig. 14.7 $E_- > E_+$). The value of the resultant of these forces is:

$$F = F_- - F_+ = qE_- - qE_+ = q(E_- - E_+). \qquad (14.22)$$

Let us introduce a ratio $(E_- - E_+)/l$, characterizing the average change in intensity per unit length of the dipole. Since the shoulder is usually small, it may be approximated as

$$(E_- - E_+)/l = \mathrm{d}E/\mathrm{d}x, \qquad (14.23)$$

where $\mathrm{d}E/\mathrm{d}x$ is a derivative of the electric field intensity along the direction of the OX axis. The derivative is a measure of heterogeneity of the electric field along the corresponding direction. From (14.23), it follows that:

$$E_- - E_+ = l\frac{\mathrm{d}E}{\mathrm{d}x},$$

then formula (14.22) can be represented as

$$F = ql\frac{\mathrm{d}E}{\mathrm{d}x} = p\cdot\frac{\mathrm{d}E}{\mathrm{d}x}. \tag{14.24}$$

Thus, a force that depends on its electric moment and the degree of nonuniformity of field $\mathrm{d}E/\mathrm{d}x$ affects the dipole.

If a dipole is oriented in a heterogenous electric field not along its lines of force, then additional torque acts on it, so a free dipole is usually pulled into the area where values of the field intensity are greater.

So far, we considered a dipole placed in an electric field, however the dipole itself is the source of the field. On the basis of (14.18) let us write down an expression for the electric potential of the field generated by a dipole at some point A at a distances from r and r_1 from the charges, respectively (fig. 14.8):

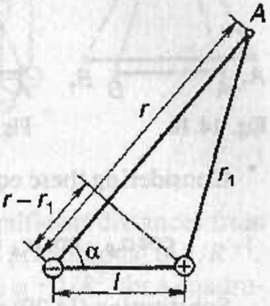


Fig. 14.8

$$\varphi = \frac{q}{4\pi\varepsilon_r\varepsilon_0}\left(\frac{1}{r_1} - \frac{1}{r}\right) = \frac{q}{4\pi\varepsilon_r\varepsilon_0}\frac{r - r_1}{rr_1}. \tag{14.25}$$

Let us assume that $l \ll r$, $l \ll r_1$, then $r \approx r_1$ and

$$rr_1 \approx r^2,\ r - r_1 \approx l\cos\alpha, \tag{14.26}$$

where α is the angle between vector p and the direction from the dipole to point A (fig. 14.8). Using (14.26), from (14.25) we obtain:

$$\varphi = \frac{ql\cos\alpha}{4\pi\varepsilon_r\varepsilon_0 r^2} = \frac{1}{4\pi\varepsilon_r\varepsilon_0}\frac{p\cos\alpha}{r^2}. \tag{14.27}$$

Let us consider some applications of formula (14.27).

Let the dipole with electric moment $\vec{P}$ be located at point O (fig. 14.9), and its shoulder be small. Using (14.27), let us write down the difference of potentials at points A and B, which are equidistant from the dipole:

$$\varphi_B - \varphi_A = \frac{1}{4\pi\varepsilon_r\varepsilon_0}\frac{p}{r^2}(\cos\alpha_B - \cos\alpha_A). \tag{14.28}$$

The angle between p and straight line AB or OC we denote as α ($\angle AOB = \beta$), then (fig. 14.9):

$$\alpha_A = \alpha + \beta/2 + \pi/2,\quad \alpha_B = \alpha - \beta/2 + \pi/2.$$

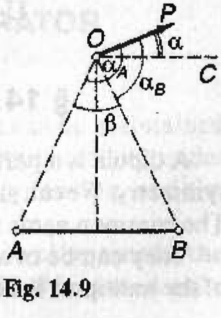


Fig. 14.9

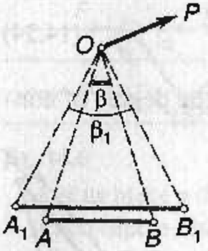


Fig. 14.10

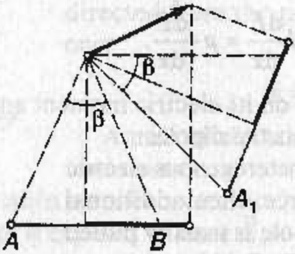


Fig. 14.11

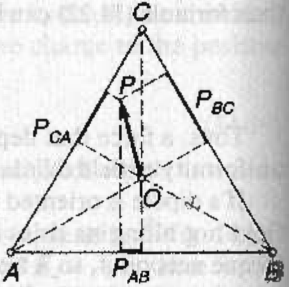


Fig. 14.12

Considering these equations, let us do trigonometric transformations:

$$\cos\alpha_B - \cos\alpha_A = -2\sin\frac{2\alpha+\pi}{2}\sin\left(-\frac{\beta}{2}\right) = 2\sin\frac{\beta}{2}\cos\alpha. \qquad (14.29)$$

Substituting (14.29) and (14.28), we have:

$$\varphi_B - \varphi_A = \frac{1}{4\pi\varepsilon_r\varepsilon_0 r^2}\frac{p}{r^2}2\sin\frac{\beta}{2}\cos\alpha = \frac{\sin\left(\frac{\beta}{2}\right)}{4\pi\varepsilon_r\varepsilon_0 r^2}p\cos\alpha. \qquad (14.30)$$

As it can be seen from (14.30), the difference of potentials between the two points of the dipole field (with ε_r and r given) depends on the sine of the half angle, at which these points are seen from the dipole (fig. 14.10), and the projection of the electric moment $p\cos\alpha$ of the dipole onto the straight line connecting the points (fig. 14.11). These observations are valid taking in account the limitations made for derivation of formula (14.27).

Let the dipole generating the electric field be in the center of an equilateral triangle ABC (fig. 14.12). Then, based on (14.30), we can graphically obtain that the intensities on the sides of this triangle are equal to the ratios of projections of p onto its sides:

$$U_{AB} : U_{BC} : U_{CA} = p_{AB} : p_{BC} : p_{CA}. \qquad (14.31)$$

§ 14.3. CONCEPT OF A MULTIPOLE

A dipole is a particular case of a system of electric charges having certain symmetry. We can give more examples of symmetric charge systems (fig. 14.13). The common name for such arrangement of charges is **electric** *multipoles.*

They can be of a different order ($l = 0, 1, 2$, etc.), the number of charges of the multipole is determined by the expression of 2l.

Thus, ***the multipole of zero order*** ($2^0 = 1$) is a single point charge (fig. 14.13, a); ***the multipole of the first order*** ($2^1 = 2$) is a dipole; ***the multipole of the second order*** ($2^2 = 4$) is a quadrupole (fig. 14.13, b), ***the multipole of the third order*** ($2^3 = 8$) is an octupole (fig. 14.13, c), etc.

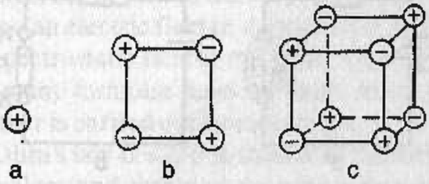


Fig. 14.13

The potential of a multipole field decreases at significant distances from the multipole ($R \gg d$, where d is its dimensions) and is proportional to $1/R^{l+1}$.

Thus, for a charge ($l = 0$) $\varphi \sim 1/R$, for a dipole ($l = 1$) $\varphi \sim 1/R^2$, for a quadrupole ($l = 2$) $\varphi \sim 1/R^3$, etc.

If a charge is distributed in a certain area of space, the potential of the electric field outside the system of charges can be represented as an approximate series:

$$\varphi = \frac{f_1}{R} + \frac{f_2}{R^2} + \frac{f_3}{R^3} + \dots \qquad (14.32)$$

In this, R is the distance from the charge system to point A with potential φ; f_1, f_2, f_3,... are some functions depending on the multipole type, its charges and the direction to point A. In (14.32), the first summand corresponds to a monopole, the second does to a dipole, the third does to quadrupole, etc. In case of a neutral charge system, the first summand is zero. If R is very large, then all members of the series starting from the third can be omitted. Then, from (14.32), we obtain the potential of a dipole [see 14.27)].

§ 14.4. DIPOLE ELECTRIC GENERATOR (CURRENT DIPOLE)

In the vacuum or an ideal insulator, an electric dipole can be maintained eternally. However, in a real situation (electrically conductive medium) under action of the electric field of the dipole, free charges are moving, and the dipole is either screened or neutralized.

It is possible to connect a voltage source to a dipole; in other words, the voltage source terminals can be represented as a dipole. In this case, despite the presence of a current in a conductive medium, the dipole will persist (fig. 14.14, a). Resistor R_1 is an equivalent of the resistance of conducting media, $\mathcal{E}$ is source emf, r is its internal resistance (fig. 14.14, b).

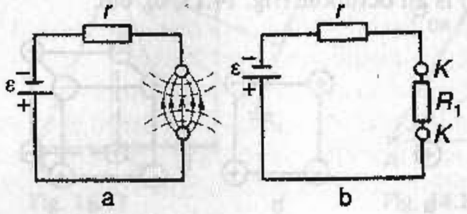


Fig. 14.14

Based on Ohm's law for a complete circuit,

$$I = \frac{\mathcal{E}}{r + R_1},$$

if $r \gg R_1$, then $I = \mathcal{E}/r$.

We can conclude that in this case, the current in the external circuit will be almost constant, it is almost independent from the properties of the media (under condition $r \gg R_1$). Such a two-pole system consisting of the source and current sink is called ***dipole electric generator***, or ***current dipole***.

There is a great analogy between dipole electric generator and electric dipole, which is based on the general analogy of the electric field in a conductive medium of the electrostatic field.

Let us illustrate this analogy with an example of a flat capacitor. Between the plates of a flat capacitor, let there be a medium with specific electrical resistance p (in other words with specific electric conductivity γ ($\gamma = 1/p$). The conductor has cross-section S and is l long, the resistance between the capacitor plates is equal to:

$$R = p\frac{l}{S} = \frac{1}{\gamma}\frac{l}{S}.$$

Electric conductivity is equal to:

$$G = \frac{1}{R} = \gamma\frac{S}{l}. \qquad (14.33)$$

If we compare (14.33) with the expression for the capacitance of a flat capacitor:

$$C = \frac{\varepsilon_r \varepsilon_0 S}{l} = \frac{\varepsilon_a S}{l}, \qquad (14.34)$$

then we can conclude: for conductivity, formula (14.33) is obtained from formula (14.34) for the capacity by replacing absolute dielectric permeability $\varepsilon_a = \varepsilon_r\varepsilon_0$ with γ.

The essence of the analogy of the electric field in a conductive medium and electrostatic field is condensed to the following:

- the current lines (an electric field in a conductive medium) coincide with the lines of an electrostatic field at the same shape of electrodes;
- in both cases, many formulas have the same form; transition from one formula to another is carried out by replacing ε_a with γ, q with I, C with G (or $1/C$ by R). Ohm's law $G = I/U$ is similar to the formula $C = q/U$.

Let us use this analogy and obtain an expression for a current dipole. Similarly to the electric moment of a dipole, we introduce the dipole moment of a dipole electric generator:

$$p_T = Il,$$

where l is the distance between current source and sink points. The potential of a dipole electric generator field is expressed by a formula which is similar to (14.27):

$$\varphi = \frac{1}{4\pi\gamma}\frac{p_T \cos a}{r^2} \tag{14.35}$$

(in an infinite medium). The intensity lines of the electrostatic field of an electric dipole are the same as the intensity lines of electric field of a current dipole (they coincide with the current lines too) (see fig. 14.3). In accordance with the statements of § 14.3, a concept of a multipole electric generator can be introduced.

Essentially, an electric multipole generator is some spatial totality of electric currents (the totality of sources and sinks of different currents).

All that was said above of potentials of charge system fields (electrostatic field), is also valid for such a generator (current multipole) in a low conducting medium.

§ 14.5. PHYSICAL PRINCIPLES OF ELECTROCARDIOGRAPHY

Living tissues are a source of electrical potentials (biopotentials).

Recording of biopotentials of tissues and organs for diagnostic (research) purpose is called ***electrography***. This general term is not very common, specific names of corresponding diagnostic methods are more popular: ***electrocardiography*** (ECG) is registration of biopotentials occurring in the heart muscle upon its excitation; ***electromyography*** is a method of registering muscular bioelectric activity; ***electroencephalography*** (EEG) is a method of registering of cerebral bioelectric activity, etc.

In most cases, biopotentials are not read with electrodes from an organ (heart, brain) directly, but it is done from other "adjacent" tissues, in which this organ generates electric fields. Clinically, this fact facilitates a registration procedure itself, making it safe and simple.

Physical approach to electrography is consists in developing (choosing) a model of electric generator. The model should correspond to the clinical picture of potentials to be read. In this connection, there are two fundamental theoretical problems: quantifying the potential in the area of measurement by the given characteristics of the electric generator (model) is a direct problem; quantifying the characteristics of the electric generator by the measured potential is an inverse problem.

Further specific consideration of physical problems of electrography will be done with the example of electrocardiography.

One of the main problems of theoretical electrocardiography is calculating distribution of cellular transmembrane potential of the heart muscle by potentials measured outside the heart. However, even theoretically, such a problem cannot be solved since the same external manifestation of cardiac biopotentials will be noted at their different internal distribution.

The physical (biophysical) approach to establishing the relationship between cardiac biopotentials and their external manifestation is simulation of sources of these biopotentials. In electrical sense, the whole heart is a sort of equivalent electric generator, either purely speculatively (hypothetically), or in the form of a real device as a totality of electrical sources in a conductor having the shape of a human body. On the surface of the conductor, when the equivalent electric generator functions, voltage is generated, which occurs on the surface of a human body due to cardiac activity. It is assumed that the medium surrounding the heart is infinite and homogenous with specific electric conductivity γ.

In this case, for a potential at some point, the formula can be written like (14.32). For large R, taking in account the assumptions made in § 14.3 in this case, we can limit ourselves to the dipole approximation and obtain formula (14.35).

Thus, we have found the expression of the potential of a dipole field. It means that in a multipole equivalent cardiac generator, on the surface of a human body, the main part of the body potential is contributed by its dipole component. In other words, it is quite accessible to simulate cardiac electrical activity if we use a dipole equivalent electric generator. If the environment is limited (finite), it is possible to come to an expression that will differ from (14.32) only by a certain multiplier.

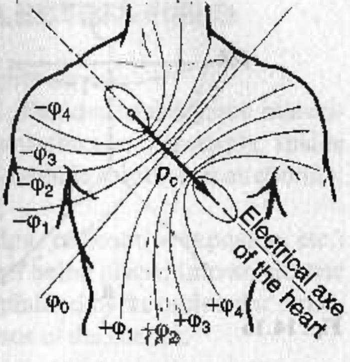


Fig. 14.15

The dipole idea of a heart underlies ***Einthoven's theory of leads***. According to it, **the heart is a dipole with dipole moment p_c[1], which rotates changing its position and point of application (change in the application point of this vector is often omitted) during a cardiac cycle.**

Fig. 14.15 shows the position of vector p_c and equipotential lines for the time point when the dipole moment is maximum; this corresponds to *R* wave on an electrocardiogram (see fig. 14.17).

For humans and some animals, the values of the maximum dipole moment are given in Table 14.1; they are compared with the masses of their heart and body.

Table 14.1

Object	**Heart mass, g**	**Body weight, kg**	**Maximum heart dipole moment, mA · cm**
Frog	0.16	0.036	0.005
Rat	1.10	0.277	0.107
Dog	108	14.2	1.63
Human	300	71.5	2.32
Horse	3060	419	13.0

W. Einthoven proposed to read the differences of cardiac biopotentials between the vertices of an equilateral triangle, which are approximately located in the right arm (*RA*), left arm (*LA*), and left leg (*LL*) (fig. 4.16, a). Fig. 14.16 gives a schematic representation of this triangle.

In the terminology of physiologists, the difference of biopotentials registered at two points of the body is called a *lead*.

The following leads are distinguished: lead I (right arm–left arm), lead II (right arm–left leg), and lead III (left arm–left leg), corresponding to differences of potentials U_{I}, U_{II} and U_{III}. According to Einthoven, the heart

---

[1] In medical and biological literature, the term "vector of the electromotive force of the heart" is used.

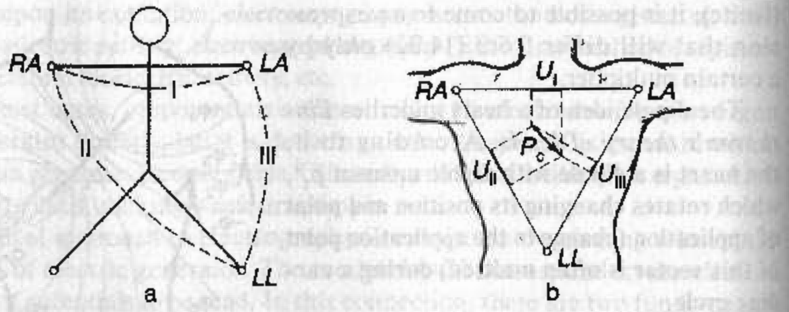


Fig. 14.16

is located in the geometric center of the triangle. The leads make it possibl to determine the ratio between the projections of the electric moment of th heart onto the sides of the triangle by formula (14.31).

Since the electrical moment of the dipole, which is the heart, changes in the course of time, then the leads will show time dependences on voltage which are called ***electrocardiograms***.

Fig. 14.17 shows a normal human electrocardiogram in one of the leads.

An electrocardiogram does not give an idea of spatial orientation of vec tor $\vec{p}_c$. However, this information is important for diagnostic purposes. In this connection, we use a method called vector-cardiography of spatial study of the cardiac electric field.

*The **vectorcardiogram** is a geometric place of points corresponding to the end of vector $\vec{p}_c$, whose position changes during the cardiac cycle.*

A projection of the vectorcardiogram onto a plane, for instance, on the frontal one, can be practically obtained summing up the voltages of two mutually perpendicular leads. Fig. 14.18 shows such a summation with the hel of an electronic oscillograph on whose screen we can see curve B. The shap of this curve is used to make diagnostic conclusions.

Considerable work on simulation of cardiac electric activity was done by L.I. Titomir.

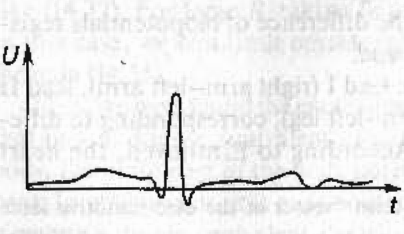


Fig. 14.17

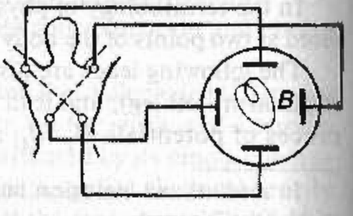


Fig. 14.18

§ 14.6. DIELECTRICS IN ELECTRIC FIELD

***Dielectrics** are bodies that do not conduct electric current.*

The term "dielectric" was introduced by M. Faraday to designate materials permeable to the electric fields, in contradistinction to metals, inside which there is no electrostatic field. Dielectrics include solids such as ebonite, porcelain, liquids like pure water, and gases.

When the external conditions change (heating, radioactive exposure, etc.) a dielectric can conduct an electric current. After being placed into an electric field, a change in the dielectric state can be explained by its molecular structure. Conventionally, let us point out three classes of dielectrics:

1) with polar molecules;
2) with non-polar molecules;
3) crystal ones.

The first class includes such materials as water, nitrobenzene, etc. Molecules of these dielectrics are not symmetrical, mass centers of their positive and negative charges do not coincide, and they have the electric moment of a dipole even in the absence of electric field.

Fig. 14.19 gives a schematic representation of molecules of hydrochloric acid (a) and water (b) as well as corresponding dipole moments in debyes[1]. In absence of an electric field, electric dipole moments of molecules are orientated chaotically (fig. 14.20, a), and the vector sum of n moments of all n molecules is equal to zero: $\sum_{i=1}^{n} \vec{p}_i = 0$.

If the dielectric is placed in an electric field, the dipole moments of molecules tend to orient themselves along the direction of the field (fig. 14.20, b); however, complete orientation cannot be attained due to molecular thermal chaotic motion. In this case $\sum_{i=1}^{n} \vec{p}_i \neq 0$.

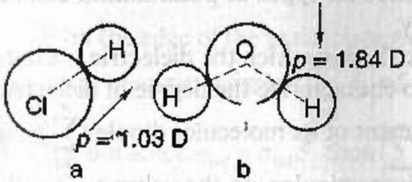


Fig. 14.19

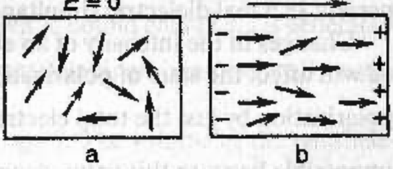


Fig. 14.20

---

[1] Debye (D) is a non-system unit of electric dipole moments of molecules: 1 D = = $3.33564 \cdot 10^{-30}$ C·m.

The second class of dielectrics includes such materials as hydrogen, oxygen, etc. whose molecules do not have dipole moments in absence of an electric field. In such molecules, the charges of electrons and nuclei are arrange in such a way that the centers of mass of positive and negative charges coincide. If a non-polar molecule is placed in an electric field, unlike charges wil slightly shift in opposite directions, and the molecule will have a dipole moment. Fig. 14.21 gives a schematic representation of molecules of such dielectric in the form of circumferences in absence (a) and presence (b) of electri field. In the circumferences, arrows indicate molecular dipole moments.

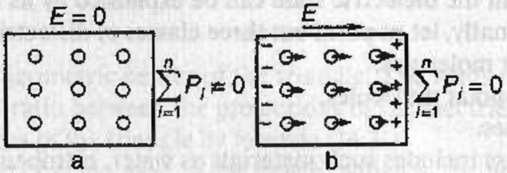


Fig. 14.21

The third class is crystalline dielectrics, like NaCl, whose lattice consis of positive and negative ions. Such dielectrics can be conventionally considered as a totality of two sublattices, one of which is charged positively, the other is charged negatively. In absence of an electric field, the sublattices are arranged symmetrically, and the total electric moment of such a dielectric is equal to zero[1]. If the dielectric is placed in an electric field, then its sublattic will slightly shift in the opposite directions, and the dielectric will acquire an electric moment.

In different dielectrics located in an electric field, all these processes are covered by the term *polarization*, that is acquiring dielectric polarity.

The first class of dielectrics is characterized by *orientational* polarization the second one — by *electronic* polarization, that is displacement of most electrons; and the third one is characterized by *ionic* polarization.

This classification is conventional since all types of polarization can be present in a real dielectric simultaneously.

Changes in the intensity of an electric field in which the dielectric is loca ed will affect the state of polarization. To characterize the degree of dielectri polarization by just the total electric moment of its molecules dipole ($\sum_{i=1}^{n} \vec{p}_i$) is impossible because this value depends, in particular, on the volume. To estimate the polarization state of a dielectric, a value called ***degree of polarizatio*** is introduced; its average value is equal to the ratio of the total electric moment of the dielectric element to the volume of this element:

---

[1] Strictly speaking, ionic crystals can have an electric moment in absence of an external field, but that is not taken into account in this.

$$\vec{P}_e = \sum_{i=1}^{n} \frac{\overline{P}_i}{V}. \tag{14.36}$$

The unit of polarization is a *Coulomb* (C) *per square meter* (C/m^2).

When a dielectric is polarized, positive charges are generated on one surface (facet) while negative charges appear on the other (see fig. 14.20, b, and 14.21, b). These electric charges are described as *bound* since they belong to the molecules of the dielectric (or to the crystal lattice in case of ion polarization) and cannot move in isolation from the molecules or be removed from the surface of the dielectric in contradistinction to free charges, which do not exist in an ideal dielectric.

As the electric field intensity increases, orientation of the molecules (orientational polarization) is ordered, and the dipole moments of the molecules increase (electronic polarization), and there is a displacement of the sublattices (ionic polarization); all this generates an increase in surface density σ_{bd} of bound electric charges.

Thus, σ_{bd} also characterizes the degree of polarization of a dielectric.

Let us determine the connection between P_e and σ_{bd} with the example of a polarized dielectric, made in the form of a parallelepiped (fig. 14.22, a). Such a parallelepiped is a totality of dipoles, which, for simplicity, can be considered as "chains"; one of them is shown in fig. 14.22, b. As the internal parts of the "chain" of dipoles are electrically compensated, then such a "chain" is similar to a long dipole with the distance between the charges equal to the edge of the parallelepiped.

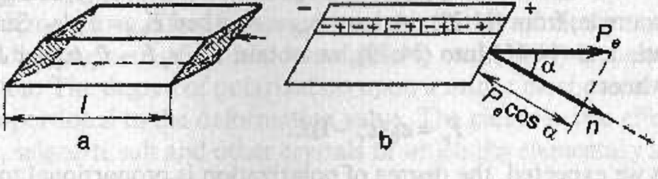


Fig. 14.22

If on the edge of the parallelepiped of area S, bound charge q_{bd} is generated, the total electric moment $\left|\sum_{i=1}^{n} \vec{p}_i\right|$ of the parallelepiped is numerically equal to $q_{\mathrm{bd}} l$, but since $q_{\mathrm{bd}} = \sigma_{\mathrm{bd}} S$, then $\left|\sum_{i=1}^{n} \vec{p}_i\right| = \sigma_{\mathrm{bd}} Sl$. The volume of the parallelepiped is $V = Sl\cos\alpha$. Based on the two last equalities, we have

$$\left|\sum_{i=1}^{n} \vec{p}_i\right| = \frac{\sigma_{\mathrm{bd}} V}{\cos\alpha}. \tag{14.37}$$

Taking into account (14.36) and (14.37), we obtain $P_e = \left|\sum_{i=1}^{n}\vec{p}_i\right|/V = \frac{\sigma_{bd}}{\cos\alpha}$, whence

$$\sigma_{bd} = P_e\cos\alpha. \qquad (14.38)$$

Thus, surface density of bound charges is equal to a component of vector $\vec{P}_e$, which is normal to the facet.

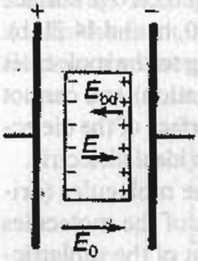


Fig. 14.23

Let us consider, for instance, a flat dielectric placed in a homogeneous electric field (fig. 14.23). E_0 is the intensity of the electric field in absence of a dielectric (field in the vacuum). Bound charges generate a homogenous field with intensity E_{bd}; as a result, there will be the electric field in the dielectric, the intensity of which is

$$E = E_0 - E_{bd}. \qquad (14.39)$$

It is known that a *relative dielectric permeability of a medium* is equal to the ratio of the interacting force of charges in vacuum to the force of interaction of the same charges at the same distance in the medium:

$$F_0/F = \varepsilon_r \text{ or } F_0 = \varepsilon_r F.$$

Since an electric field intensity is proportional to the force acting on a charge [see (14.1)], a similar ratio can be written for E_0 and E:

$$E_0 = \varepsilon_r E. \qquad (14.40)$$

The electric field intensity generated by bound electric charges $E_{bd} = \sigma_{bd}/\varepsilon_0$. For this example, from (14.39), we have $\sigma_{bd} = P_e$, then $E_{bd} = P_e/\varepsilon_0$. Substituting this formula and (14.40) into (14.39), we obtain $E = \varepsilon_r E - P_e/\varepsilon_0$, or $E(\varepsilon_r - 1) = P_e/\varepsilon_0$, whence,

$$P_e = \varepsilon_0(\varepsilon_r - 1)E. \qquad (14.41)$$

Just as we expected, the degree of polarization is proportional to the electric field intensity in a dielectric. Based on (14.41), we introduce a concept of *dielectric susceptibility of a medium*:

$$\chi = \varepsilon_r - 1, \qquad (14.42)$$

which characterizes the ability of a dielectric to polarize together with relative dielectric permeability ε_r and depends on its molecular structure of the and on its temperature. If the electric field is variable, ε_r and χ change depending on the frequency.

Table 14.2 shows values of ε_r of various biological environments and some materials in a constant electric field at room temperature.

Table 14.2

Material	ε_r	Material	ε_r
Kerosene	2	Egg white	72
Vegetable oil	2–4	Water	81
Glass	6–10	Whole blood	85
Starch	12	Gray cerebral material	85
Cow milk	66	Optic nerve	89
		White cerebral matter	90

Attempts are made nowadays to use differences in dielectric permeability of normal and pathological tissues and media in both constant and variable electric fields for diagnostic purposes.

§ 14.7. PIEZOELECTRIC EFFECT

In crystalline dielectrics, polarization can occur in absence of an electric field upon deformation. This phenomenon is called ***piezoelectric effect*** (*piezoeffect*).

They distinguish *transverse* piezoelectric effect (fig. 14.24) and *longitudinal* one (fig. 14.25). The arrows show forces acting on the crystal. When deformation changes, for instance upon transition from compression to stretching, the sign of the arising polarization charges will also change.

The piezoelectric effect is caused by deformation of elementary crystal cells and displacement of sublattices relative to each other upon mechanical deformation. The degree of polarization upon a minor mechanical deformation is proportional to the deformation value. The piezoelectric effect occurs in quartz, seignette salt and other crystals in which the elementary lattice cell has no the center of symmetry.

To demonstrate a piezoelectric effect, it is possible to use an installation, whose diagram which is shown in fig. 14.26. To crystal *C* with piezoelectric

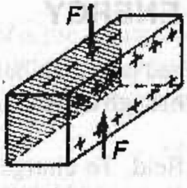


Fig. 14.24

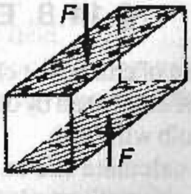


Fig. 14.25

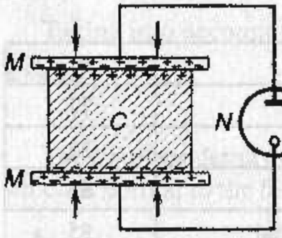


Fig. 14.26

properties, metal plates M are applied, which are connected through neon lamp N. This lam consumes a low intensity current and lights up at a certain voltage, that is it is a kind of voltag tester.

An impact (deformation) of the crystal gen- erates a voltage on its facets, and, therefor on metal plates too, after that the neon lam flashes.

Along with the direct piezoelectric effect considered, a reverse piezoele tric effect is observed too: when an electric field is applied to crystals, they are deformed.

Both direct and reverse piezoelectric effects are used when it is necessa to convert a mechanical quantity into an electrical one, or vice versa.

Thus direct piezoelectric effect is used in medicine: in sensors, for puls recording; in technology it is used in adapters and microphones and for mea suring vibrations; and reverse piezoelectric effect is used to produce mechani cal oscillations and waves of ultrasonic frequency.

A sizeable piezoelectric effect occurs in bone tissue in case of shear defor mations.

The cause of the effect is deformation of collagen, which is the main connective tissue protein. Therefore, tendons and skin also have piezoelectri properties. Under normal functional loads, as well as in absence of defects bone structure, only compression and stretching deformation exist in it, b no piezoelectric effect. When something is abnormal and there is shear defor mation, a piezoelectric effect occurs. It affects the continuous processes of destruction and regeneration in the bone and helps to ensure that the shea vanishes (changing the architecture and even shape of bone). Two possib mechanisms of the piezoelectric effect are observed: a) the electric fie changes the activity of cells producing collagen, and b) the electric field is involved into packing of macromolecules. This problem was studied by V.F. Chepel.

§ 14.8. ELECTRIC FIELD ENERGY

A system of charges or charged bodies and charged capacitor has an energy.

It can be evidenced by discharging a capacitor through a light bulb attached to it: the bulb will flash.

Let us calculate the energy of the capacitor field. To charge it, we wil transfer positive charge dq from one plate to the other repeatedly. In the process of transfer, the voltage between the capacitor plates increases. The work to be done against forces of the electric field to charge the capacitor is equal to the ***energy of the capacitor***:

$$E_{el} = A.$$

The elementary work moving a charge against the forces of an electric field is $dA = U \cdot dq$. Transfer of charge dq from one capacitor plate to the other changes voltage dU of the capacitor, and then based on the formula for electrical capacity, we write down $dq = C \cdot dU$, and it means that $dA = C \cdot U \cdot dU$.

Integrating this equation in the limits from $U_0 = 0$ to some final value U, we find an expression for the ***energy of electric field of a charged capacitor***:

$$A = E_{el} = C \cdot U^2/2. \qquad (14.43)$$

Or, taking into account that $C = q/U$,

$$E_{el} = q \cdot U/2 = q^2/(2C). \qquad (14.44)$$

If, without changing the charge on the plates of a capacitor disconnected from a voltage source, we pull apart its plates from l_1 to l_2, then the electric capacity will decrease. As it can be seen from (14.44), as the volume occupied by the electric field increases, the energy of the capacitor (fig. 14.27), will increase too, but the field intensity is constant. Hence, it is clear that the energy of the charged capacitor is concentrated in the volume occupied by the electric field.

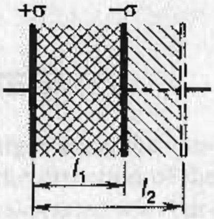


Fig. 14.27

Existence of the energy of an electric field can be explained more convincingly with an example of an alternating electromagnetic field (transfer of a signal over a long distance, light pressure, etc.).

Let us express the energy of the field through its characteristics. For this purpose, let us transform (14.43) by substituting the expression for the capacitance of a flat capacitor and voltage from (14.14):

$$E_{el} = \varepsilon_r \varepsilon_0 E^2 V/2, \qquad (14.45)$$

where $V = Sl$ is the volume occupied by the electric field.

Assuming that the electric field of a flat capacitor is homogeneous, we divide (14.45) by the volume and obtain ***volumetric energy density of the field***:

$$\omega_{el} = \varepsilon_r \varepsilon_0 E^2/2. \qquad (14.46)$$

The unit of volumetric density is a *Joule per cubic meter* (J/m^3).

In conclusion, we note that formula (14.46) is also valid for a non-unifor electric field, but then it expresses the volume density of energy at a poin The energy of this field can be found by integrating (14.46) for the corresponding volume:

$$E_{el}=\frac{\varepsilon_0}{2}\int\limits_V \varepsilon_r E^2\,\mathrm{d}V.$$

In general, relative dielectric permeability varies at different points of the medium, that is it depends on the coordinates, so in this formula, ε_r is under the sign of the integral.

Chapter 15

Electric current

An electric current is usually understood as directed flow of electric charges. Conduction current and convection current are distinguished. The former is directed flow of charges in conducting bodies: electrons in metals, electrons, as well as holes, in semiconductors, ions in electrolytes, ions, as well as electrons in gases. Convection current is flow of charged bodies and a flow of electrons or other charged particles in a vacuum.

The given current classification is conventional to a great extent. Thus, an alternating electric field is also called a current — displacement current. There is at least one common indicator of any current: it is a source of magnetic field.

In this chapter we consider some characteristics of electric current and current sources, current in electrolytes and gases, and thermoelectric phenomena.

§ 15.1. CURRENT DENSITY AND INTENSITY

The trajectories of directed flow of positive electric charges along the conductor are called *current paths*, tangents to which show the direction of the velocity of ordered charge flow. Current paths are usually associated with current density rather than with the velocity of charge.

Current density is a vector characteristic of an electric current, numerically equal to the ratio of the intensity of a current passing through a small surface element, normal to the direction of motion of charged particles, forming the current, to the area of this element:

$$j = \mathrm{d}I/\mathrm{d}S.$$

In § 13.4, a relationship between the particle flux density, concentration and velocity of their directional motion was found [see (13.29)]:

$$J = n\upsilon.$$

If we multiply this formula by charge q of the current carrier, we obtain the current density:

$$j = qJ = qn\upsilon. \tag{15.1}$$

It can be given also in the vector form:

$$\vec{j} = qn\vec{\upsilon}. \tag{15.2}$$

Vector $\vec{j}$ is tangential to the current paths. For current intensity, let us write down the following expression:

$$I = \mathrm{d}q/\mathrm{d}t. \tag{15.3}$$

Current intensity is a time derivative of the charge passing through some section or surface.

§ 15.2. ELECTROMOTIVE FORCE OF CURRENT SOURCES

In order for direct current to flow through a conductor, it is necessary to maintain a potential difference at its ends. This is carried out by current sources.

Let a positive charge move along a closed circuit (fig. 15.1). Ideally, resistance of the supply conductors (segments *1–2* and *3–4*) can be taken as zero, i.e. the potentials of points *1* and *2* (*3* and *4*) are equal. From (14.15), it follows that the field intensity in such conductors is zero. Directed motion of charges takes place by inertia, without resistance or impact of an accelerating force.

In segment *2–3* the difference of potentials ($\varphi_2 - \varphi_3$) is equal to the drop of voltage IR. Presence of such a difference means that the electric field intensity is other than zero in the conductor. Therefore, a force from the electric field affects the charge, the charges interact in metals with ions of the crystal lattice, which generates a friction force (electrical resistance).

In segment *4–1*, the positive charge passes from lower potential φ_4 to higher one φ_1. Such motion against the forces of electric field occurs under action of forces called ***external*** ones (F_{ext}). These forces can be chemical, electromagnetic, mechanical and other except electrostatic ones.

The external forces perform work.

An electromotive force (emf) of a current source is a value, which is numerically equal to the work of external forces moving a unit positive charge along the whole circuit.

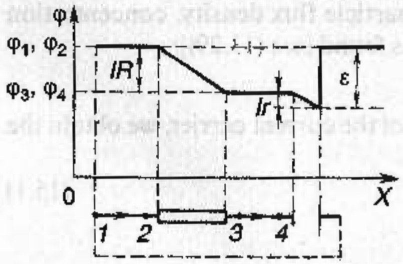


Fig. 15.1

In practice, the work of external forces is other than zero only inside a current source. According to (14.1), the ratio of an external force to a unit positive charge is equal to the field intensity of external forces:

$$\vec{E}_{\text{ext}} = \vec{E}_{\text{ext}}/q. \tag{15.4}$$

From the definition of emf and general formula for work, we can write down:

$$\mathcal{E} = \oint qE_{extl}\,dl, \tag{15.5}$$

where E_{extl} is projection of the field intensity of external forces onto direction d*l*.

In this, you can not integrate along the entire circuit, but only along the segments, in which the current sources are located. As it can be seen from (15.5), the emf in the circuit is equal to circulation of the field intensity of external forces along the circuit.

Along with the increase in potential in direction *4–1* within the source, whose resistance is *r*, there is also a potential drop equal to *Ir* (fig. 15.1). The figure represents potential distribution along the circuit shown below the plot.

Emf corresponds to an abrupt change in the potential in the current source.

§ 15.3. ELECTRICAL CONDUCTIVITY OF ELECTROLYTES

Biological liquids are electrolytes whose electrical conductivity is similar to that of metals: unlike gases, in both media, current carriers exist independently from an electric field, therefore, expression (15.1) is also valid for electrolytes; nevertheless, in contrast to metals, it should be presented separately for positive and negative ions:

$$j_+ = qn_+ \cdot \upsilon_+ \text{ and } j_- = qn_- \cdot \upsilon_-. \tag{15.6}$$

Total current density:

$$j = j_+ + j_- = q(n_+ \cdot \upsilon_+ + n_- \cdot \upsilon_-). \tag{15.7}$$

If we assume that each molecule dissociates into two ions, the concentration of positive and negative ions is the same:

$$n_+ = n_- = \alpha n, \tag{15.8}$$

where α is dissociation coefficient; n is concentration of electrolyte molecules.

Directional motion of ions in an electric field can be considered to be homogenous, while force qE acting on an ion from the electric field, is equal to friction force $r \cdot \upsilon$:

$$qE = r \cdot \upsilon,$$

whence, substituting $q/r = b$, we obtain:

$$\upsilon = bE. \tag{15.9}$$

The proportionality coefficient b is called ***mobility of ions***. It is equal to the ratio of the velocity of directional motion of ions generated by the electric field to the intensity of this field[1].

For ions of different signs, from (15.9), we have respectively:

$$\upsilon_+ = b_+ \cdot E \text{ and } \upsilon_- = b_- \cdot E. \tag{15.10}$$

Substituting (15.8) and (15.10) into (15.7), we obtain:

$$j = nq\alpha \cdot (b_+ + b_-)E. \tag{15.11}$$

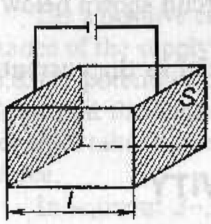


Fig. 15.2

Let us represent an electrolyte in the form of a rectangular parallelepip with electrode facets with area S located at distance l (fig. 15.2). Let us multip both parts of (15.11) by S. Taking into account (14.14), we obtain:

$$jS = nq\alpha \cdot (b_+ + b_-) \cdot (U/l)S. \tag{15.12}$$

Since $jS = I$, then (15.12) meets Ohm's law for the segment of the circuit without a current source: $I = U/R$, where:

$$R = (l/S)[nq\alpha(b_+ + b_-)]^{-1} \tag{15.13}$$

is the electrolyte resistance. Comparing (15.13) with relation $R = \rho l/S$, we obtain:

$$\gamma = 1/\rho = nq\alpha(b_+ + b_-). \tag{15.14}$$

From here it follows that specific conductivity of electrolyte is the higher, the higher the ion concentration, their charge and mobility are. Electric conductivity of electrolytes is proportional to its temperature.

§ 15.4. ELECTRICAL CONDUCTIVITY OF BIOLOGICAL TISSUES AND LIQUIDS AT DIRECT CURRENT

Biological tissues and organs are rather heterogeneous objects with diffe rent electrical resistances, which can change under the effect of an electri current. Such a circumstance causes difficulties in measuring electrical resis tance of living biological systems.

---

[1] Mobility b is connected to mobility u (see § 13.3) by relation $b = uq$.

The electrical conductivity of individual parts of a body located between electrodes applied onto a surface of a body directly, depends on the resistance of the skin and subcutaneous layers very much.

Inside the body, a current spreads firstly along the blood and lymph vessels, muscles, sheaths of nerve trunks. The skin resistance, in turn, is determined by its state: thickness, age, moisture, etc.

The electrical conductivity of tissues and organs depends on their functional state and, therefore, can be used as a diagnostic index. Thus, in inflammation, when cells swell, the cross-section of intercellular joints decreases and electrical resistance increases too; physiological phenomena that cause perspiration are accompanied by an increase in the electrical conductivity of skin, etc.

Below we present specific resistance of various tissues and body liquids (Table 15.1). Electrical conductivity of tissues at an alternating current (AC) will be considered in § 18.4.

Table 15.1

Material	ρ, **Ohm·m**
Cerebrospinal fluid	0.55
Blood	1.66
Muscles	2
Cerebral and nerve tissue	14.3
Fatty tissue	33.3
Dry skin	10^5
Bone without periosteum	10^7

§ 15.5. ELECTRIC DISCHARGE IN GASES. AIR IONS AND THEIR THERAPEUTIC AND PREVENTIVE EFFECT

A gas consisting only of neutral particles is an insulator. If it is ionized, it becomes electrically conductive. *Any device, phenomenon, factor that can ionize gas molecules and atoms, is called an ionizer.*

This could be visible light, X-rays, a flame, ionizing radiation, etc. An electrical charge in the air can be generated by spraying polar liquids in the air (balloelectric effect), i.e. such liquids whose molecules have a constant electric dipole moment. For instance, for dispersion in the air, water breaks down into charged droplets. The charge sign of large droplets (positive for pure water) is opposite to the sign of the smallest droplets. Greater drople settle down quite quickly, and negatively charged water particles remain in the air. This phenomenon is observed in a fountain.

Electric conductivity of gas also depends on secondary ionization.

To ionize a neutral atom, it is necessary to do some work A_i for an electro detachment; such work is equal the energy of ionization. In physics, it is ac energy (work) of ionization is conventionally expressed through ***ionizatio potential***:

$$\varphi_i = A_i/e. \qquad (15.15)$$

Thus, in volts, an ionization potential is numerically equal to ionizatio work ionization in electron-volts.

Let us give the values of the lowest ionization potential for some gases co responding to detachment of outer shell electrons (Table 15.2).

Table 15.2

Material	φ_i, V	Material	φ_i, V
Sodium vapors	5.1	Nitrogen	15.5
Mercury vapors	10.4	Hydrogen	15.6
Oxygen	12.5	Helium	21.5
Carbon dioxide	14.4		

The ionization potential of internal electrons is much higher.

Along with ionization, there is also a reverse process, i.e., recombination ions, at which energy is released. An example of this phenomenon is glowing gas discharge tubes. If an ionizer does not function, then due to recombinatio in absence of an electric field, the gas will become an insulator quite quickly.

In terrestrial conditions, the air usually contains a certain amount of ion thanks to natural ionizers; firstly, they are radioactive materials in soil an gases, space radiation. In the air, ions and electrons associate themselves with neutral molecules and suspended particles and form ions that are more com plex. These ions in the atmosphere are called ***aeroions***.

They differ not only in their sign, but also in mass; they are conventiona divided into light (gas ions) and heavy (suspended charged particles, which are specks of dust, particles of smoke and liquid).

Heavy ions are harmful to the body. Light and mostly negative aeroio have a salutary effect. They are used, in particular, for treatment purposes — ***aeroionotherapy***.

We distinguish natural aeroionotherapy when a patient is in natural conditions with enhanced air ionization (mountains, waterfalls, etc.)[1] and artificial one implemented with special devices — ***aeroionizers***. Any device producing ions in the air can be an ionizer. Nevertheless, when used for medicinal purposes, it should not cause adverse side effects on the body. A kind of artificial aeroionotherapy is an electrostatic shower (***franklinization***). In franklinization, high voltage stationary electric field is used (up to 50 kV). The therapeutic effect is produced by aeroions and a small amount of ozone being formed. Franklinization is administered in the form of general and topical treatment. In case of general franklinization, the patient sits on an insulated wooden chair with a metal plate, connected to the positive pole of the device. Above the patient's head, 10–15 cm higher, an electrode in the form of a "spider" is placed; the electrode is connected to the negative pole of the apparatus.

§ 15.6. INTERNAL CONTACT POTENTIAL. THERMOELECTROMOTIVE FORCE

Let us consider a contact of two metals *1* and *2* with different concentrations of free electrons: $n_1 > n_2$ (fig. 15.3, a). A contact between them will generate diffusion of electrons one each metal to the other. Since the electron concentrations are different, the diffusing flows from different metals will be different. This will lead to the metals being filled up with opposite charges producing an internal contact potential difference U_i between them. At that all, the first metal has a higher potential relative to the second (fig. 15.3, a). In the near-contact zone at a steady state value of contact potential difference (fig. 15.3, b), a change in energy E_e of free electrons corresponds to a dynamic equilibrium. In the state of dynamic equilibrium, electron flows in the two directions are equal. Since in metals the concentration of free electrons is very high, transfer of electrons from each metal to the other will hardly produce any change in their concentrations which will remain the same in dynamic equilibrium conditions (n_1 and n_2) as well. The internal contact potential

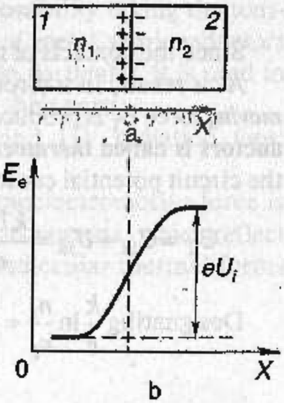


Fig. 15.3

---

[1] A change in the ionic composition of the air due to solar activity is probably one a causes of solar effect on terrestrial biological organisms. This effect is studied in a section of biophysics called *heliobiology*.

difference can be determined on the basis of the general equilibrium condition, which is equation of the electrochemical potentials of contacting metals (see § 12.6):

$$\Delta\tilde{\mu}_i = 0 \quad \text{or} \quad RT\ln\frac{n_2}{n_1} + ZF(\varphi_2 - \varphi_1) = 0.$$

Since $Z = -1$, then

$$U_i = \varphi_2 - \varphi_1 = \frac{RT}{F}\ln\frac{n_1}{n_2}. \tag{15.16}$$

Thus, the internal contact potential difference depends both on the differences in concentrations of free electrons in metals and on contact temperature.

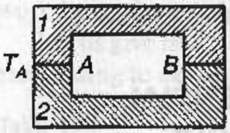


Fig. 15.4

Let us consider a closed circuit consisting of two metals *1* and *2* (fig. 15.4), in which the concentration of free electrons n_1 and n_2 respectively. On contacts A and B of metals, temperatures T_A and T_B are maintained respectively. To be specific, let us assume that $n_1 > n_2$ and $T_A > T_B$. Let us write down expression (15.16) for both contacts:

$$U_{iA} = \frac{kT_A}{e}\ln\frac{n_1}{n_2}, \tag{15.17}$$

$$U_{iB} = \frac{kT_B}{e}\ln\frac{n_2}{n_1}. \tag{15.18}$$

Since the contacts of metals have different temperatures, $U_{iA} \neq U_{iB}$.

As a result, in a circuit consisting of different metals, a ***thermoelectric moving*** **force** $\mathcal{E}_T$ is produced. **This phenomenon, which is also valid for semiconductors is called *thermoelectricity*.** Since emf is equal to the sum of jumps of the circuit potential caused by extraneous forces, then

$$\mathcal{E}_T = U_{iA} + U_{iB} = \frac{kT_A}{e}\ln\frac{n_1}{n_2} + \frac{kT_B}{e}\ln\frac{n_2}{n_1} = \frac{k}{e}\ln\frac{n_1}{n_2}(T_A - T_B). \tag{15.19}$$

Designating $\frac{k}{e}\ln\frac{n_1}{n_2} = \beta$, we obtain

$$\mathcal{E}_T = \beta(T_A - T_B). \tag{15.20}$$

The device shown in fig. 15.4 is called ***thermal element***, or a ***thermocouple***.

From (15.20), it is clear that β corresponds to thermal emf generated in a circuit at a temperature difference of the contacts equal to 1 K, and is a thermocouple characteristic.

Let us give the value of β at temperatures about 100 °C for some couples of metals (Table 15.3).

Table 15.3

Couple	β, μV/K	Couple	β, μV/K
Zn–Ag	0.5	Mg–Ag	3.5
W–Ag	2.5	Mo–Ag	6.3
Pb–Ag	3.0	Fe–Pt	18.1

A sizeable value of thermal emf is attained not only by choosing an appropriate pair of metals or semiconductors or increasing ΔT but also by serial connection of several thermocouples in a thermobattery (thermopile). Fig. 15.5 shows a thermopile consisting of four thermocouples (odd contacts *1*, *3*, *5*, *7* have one temperature, even contacts *2*, *4*, *6*, *8* have a different one).

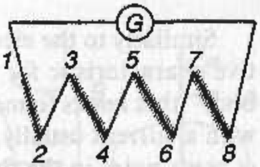


Fig. 15.5

Thermoelectricity has three main applications:

1) developing current generators with direct conversion of molecular thermal energy into electrical one: modern semiconductor thermoelectric generators have the efficiency of about 10%;
2) temperature taking: knowing the dependence $\mathcal{E}_T = f(\Delta T)$ and based on measurements of $\mathcal{E}_T$ we can find ΔT, and, therefore, T. The convenience of this method consists in remotability and possibility taking the temperature of small objects because a contact of metal semiconductors itself can be made quite small. In medicine, in particular, it is used to take the temperature of individual organs and their parts;
3) measuring the power of infrared, visible and UV radiation (see, the structure of an actinometer in § 27.4).

In the example considered, generation of thermoelectromotive force is a ***thermoelectric phenomenon***. This term describes of phenomena, which reflect a specific connection between the electrical and molecular thermal forms of impermanence in metals and semiconductors.

Chapter 16

Magnetic field

A magnetic field is a type of matter by means of which a force acts on moving electric charges placed in the field, and on other bodies with a magnetic moment. A magnetic field is one of the manifestations of the electromagnetic field.

§ 16.1. MAGNETIC FIELD INDUCTION

Similarly to the electrostatic field, it is necessary to introduce a quantitative characteristic for a magnetic field. To do this, an object called a "test body" that reacts to magnetic fields is chosen. For this purpose a small frame with a current usually suffices, so we can consider that the frame is placed at some point in the field. Experiments prove that in a magnetic field, a test frame with a current is exposed to a moment of force M; the latter depends on a number of factors including orientation of the frame. The maximum value $M_{\max}$ depends on the magnetic field, in which the contour is located, and on the contour itself: the intensity I of the current flowing along it and area S bounded by the circuit, i.e.

$$M_{\max} \sim IS. \qquad (16.1)$$

The value

$$p_m = IS \qquad (16.2)$$

is called a *magnetic moment* of a loop of current. Therefore,

$$M_{\max} \sim p_m. \qquad (16.3)$$

A magnetic moment is a vector quantity. For the a loop of current, vector $\vec{p}_m$ is directed perpendicularly to the plane of the circuit and is connected with the direction of current I by the rule of the right-hand screw (fig. 16.1).

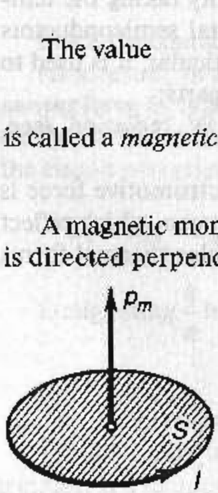


Fig. 16.1

A magnetic moment is a characteristic not only of a loop of current, but also of many elementary particles (protons, neutrons, electrons, etc.), determining their behavior in the magnetic field.

The unit of a magnetic moment is *Ampere-square meter* ($A \cdot m^2$). The magnetic moment of elementary particles, nuclei, atoms, and molecules is expressed in special units called either *atomic* (μ_B), or *nuclear* (μ_N) *Bohr magneton:*

$$\mu_B = 0.927 \cdot 10^{-23}\ \mathrm{A \cdot m^2\ (J/T)},$$
$$\mu_N = 0.505 \cdot 10^{-26}\ \mathrm{A \cdot m^2\ (J/T)}.$$

Dependence (16.3) is used to introduce a force characteristic of a magnetic field — magnetic induction B.

At a certain point of the field, magnetic induction is equal to the ratio of the maximum torque acting on a current loop in a homogeneous magnetic field to the magnetic moment of this loop:

$$\mathrm{B} = M_{\max}/p_m. \tag{16.4}$$

Vector $\vec{B}$ coincides in direction with vector $\overline{p_m}$ in the state of stable equilibrium of the circuit. Fig. 16.2 shows the position of a current loop in the magnetic field with induction B corresponding to the maximum (a) and zero (b) moment of force. The last case corresponds to stable equilibrium (vectors B and $\overline{p_m}$ are collinear).

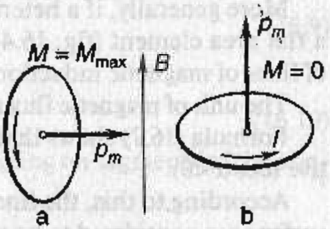


Fig. 16.2

The unit of magnetic induction is *Tesla* (T):

$$1\ \mathrm{T} = \frac{1\ \mathrm{N \cdot m}}{1\ \mathrm{A \cdot m^2}} = 1\ \mathrm{N/(A \cdot m)}.$$

Thus, in the field with magnetic induction 1 T, a loop with a magnetic moment of 1 A · m^2 is exposed to the maximum moment of force 1 N · m.

A magnetic field is graphically depicted using lines of ***magnetic induction***, tangents to which show the direction of vector B. The density of lines, i.e. the number of lines passing through a unit area element perpendicularly to them, is equal to the absolute value of vector B. Magnetic induction lines have neither the beginning nor end and are closed. Such fields are called ***vortex field***. **The circulation of a magnetic induction vector along any *magnetic induction* line is not zero:**

$$\oint = B_l\,\mathrm{d}l \neq 0. \tag{16.5}$$

Let us consider an area element S located in a region of a homogeneous magnetic field with induction B (fig. 16.3)[1]. Let us draw lines of magnetic

---

[1] Since the lines of B are closed, then the magnetic field cannot be homogenous infinitely.

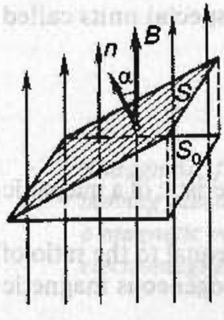


Fig. 16.3

induction through this area element. Its projection onto a plane, which is perpendicular to the lines, is S_0. The number of lines passing through S and S_0 is equal. Since the density of lines corresponds to value B, the total number of lines passing through the area element is equal to

$$\Phi = BS_0. \quad (16.6)$$

It is seen in fig. 16.3 that $S_0 = S\cos\alpha$, whence

$$\Phi = BS\cos\alpha \text{ or } \Phi = B_nS, \quad (16.7)$$

where $B_n = B\cos\alpha$ is the projection of vector B onto the direction of normal n to the area element; Φ is ***magnetic flux***.

More generally, if a heterogeneous magnetic field is a surface rather than a flat area element (fig. 16.4), magnetic flux Φ is also equal to the number of lines of magnetic induction passing through the surface.

The unit of magnetic flux according to (16.6), is *weber* (W): 1 Wb = 1 T·m^2.

Formula (16.7) shows that a flux can be both positive ($\cos\alpha > 0$) and negative ($\cos\alpha < 0$).

According to this, the lines of magnetic induction going out from a closed surface are considered to be positive, and incoming lines are considered to be negative. Since magnetic induction lines are closed, the magnetic flux passing through a closed surface is zero (fig. 16.5).

§ 16.2. AMPERE LAW. ENERGY OF A LOOP OF CURRENT IN A MAGNETIC FIELD

One of the main manifestations of a magnetic field is its force action on moving electric charges and currents. Having generalized numerous experi

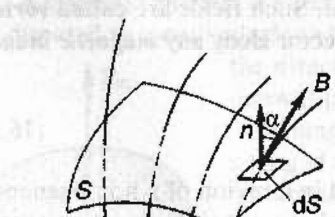


Fig. 16.4

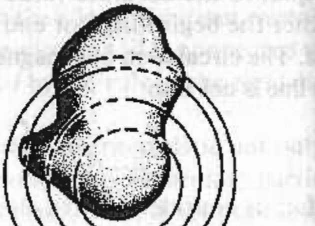


Fig. 16.5

mental data, A.M. Ampere established a law determining this force effect.

Let us reduce it to the differential form, which will permit quantifying the force acting on various loops of current located in a magnetic field. In a conductor located in a magnetic field, we select a sufficiently small segment dl, which can be considered to be a vector directed along the current (fig. 16.6). Product Idl is called a ***current element***. The force acting from the magnetic field onto the current element is:

$$dF = kIB\sin\beta \cdot dl, \qquad (16.8)$$

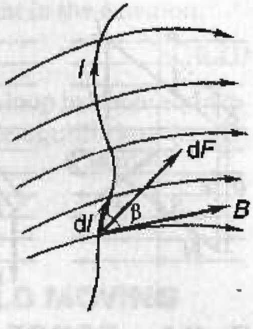


Fig. 16.6

where k is the proportionality coefficient; in the SI system, $k = 1$, that is why

$$dF = IB\sin\beta\, dl, \qquad (16.9)$$

or in a vector form:

$$d\vec{F} = Id\vec{l} \times \vec{B}. \qquad (16.10)$$

Integrating this equation, we find the force acting on segment l of the conductor from the magnetic field:

$$F = I\int_l (d\vec{l} \times \vec{B}). \qquad (16.11)$$

Formulas (16.8)–(16.10) express the **Ampere *law***.

Let us consider some examples of application of formula (16.10).

1. A rectilinear segment of a conductor with current I and length l, is located in a homogeneous magnetic field at angle β to magnetic induction B (fig. 16.7). To find the force acting on this segment of the conductor from the magnetic field, we integrate (16.11) and obtain:

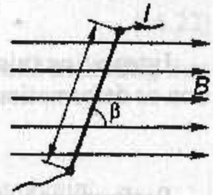


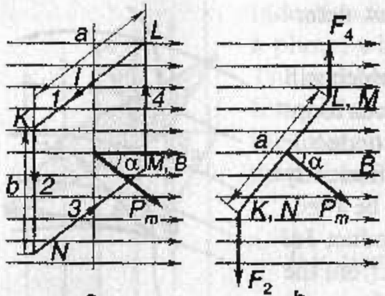
Fig. 16.7

$$F = IBl\sin\beta. \qquad (16.12)$$

2. Rectangular frame $KLMN$ with current I is located in a homogeneous magnetic field with induction B (fig. 16.8, a). Let us number the sides of the frame and designate the forces acting on them from the magnetic field as F_1, F_2, F_3, F_4.

Forces F_1 and F_3, applied to the bisecting point of corresponding sides are directed oppositely and are equal according to formula (16.12). Forces $F_2 = IBb$ and $F_4 = IBb$ generate a pair of forces, the moment of which (fig. 16.8, b)

$$M = IBb(a/2)\sin\alpha + IBb(a/2)\sin\alpha = IBba\sin\alpha. \qquad (16.13)$$



a b

Fig. 16.8

Since $Iba = IS = p_m$, then from (16.13 we obtain:

$$M = p_m B \sin\alpha. \quad (16.1$$

or, in the vector form:

$$\vec{M} = \vec{p_m} \times \vec{B}. \quad (16.15)$$

In fact, magnetic induction was introduced in § 16.1 based on this dependenc

Using the Ampere law, we calculat the work of a magnetic field when a loop of current moves or changes its shape.

In case forces F_2 and F_4 do a positive work (rotation counterclockwis fig. 16.8, b), angle α reduces ($d\alpha < 0$), therefore, for rotating the fram [see (5.13)] $dA = -M d\alpha$ or taking into account (16.14),

$$dA = -Bp_m \sin\alpha\, d\alpha = -IBS \cdot dE_p, \quad (16.16)$$

(we consider that upon an elementary rotation $d\alpha$, current intensity I does not change).

Differentiating (16.7), we obtain:

$$d\Phi = -BS \sin\alpha\, d\alpha. \quad (16.17)$$

Comparing (16.16) and (16.17), we obtain:

$$dA = I d\Phi.$$

Integrating this equation, we find the work of magnetic field forces at motion or deformation of a loop of current in the field:

$$A = \int I d\Phi. \quad (16.18)$$

Performance of work by forces of a field means a change in the energy of the loop of current. It could be a change in energy associated with motion the loop (kinetic one) or a change in its position (potential one), or both. In this case, the loop does not accelerate; therefore, only its potential energ changes in the magnetic field. Work is a measure of energy transfer from one body to the other, so elementary work of the field forces is equal to the ele mentary change in the potential energy of the loop of current in the magneti field: $dA = -dE_p$. Taking into account (16.16), we write down:

$$dE_p = p_m B \sin\alpha\, d\alpha. \quad (16.19)$$

Integrating this expression, we obtain:

$$E_p = -p_m B \cos\alpha + \text{const}. \quad (16.20)$$

From condition $E_p = 0$ at $\alpha = \pi/2$ we find the constant in the equation:

$$E_p = -p_m B\cos\alpha. \tag{16.21}$$

Formula (16.21) shows that the potential energy of a loop in stable equilibrium ($\alpha = 0$) is minimum: $E_p = -p_m B$, and in unstable equilibrium ($\alpha = \pi$) is maximal:

$$E_p = p_m B.$$

§ 16.3. EFFECT OF MAGNETIC FIELD MOVING ELECTRIC CHARGE. LORENTZ FORCE

According to the Ampere law, a force acting on a current-carrying conductor in a magnetic field is the result of its action on moving electric charges that make up this current.

Let us consider a cylindrical conductor l long with current intensity I, located in a magnetic field of induction B (fig. 16.9). The velocity of directional motion of positive charge q is equal to υ. The force acting onto a single moving charge is determined by the ratio of force F, applied to the current conductor to the total number of N carriers of current in it:

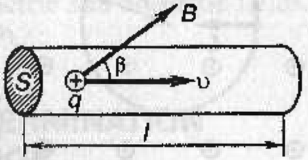


Fig. 16.9

$$f_1 = F/N. \tag{16.22}$$

Let us expand the expression for the force using (16.12) and assuming that the current is $I = jS$:

$$F = jSBl\sin\beta,$$

where j is the current density. Taking into account (15.1), we obtain:

$$F = jSBl\sin\beta = qn\upsilon SBl\sin\beta, \tag{16.23}$$

where $n = N/(Sl)$ is particle concentration. Substituting (16.23) into (16.22), we obtain an expression for a force acting from a magnetic field on a separate moving electric charge and called the ***Lorentz force***:

$$f_1 = \frac{qN\upsilon SBl\sin\beta}{SlN} = q\upsilon Bl\sin\beta. \tag{16.24}$$

The direction of the Lorentz force can be determined from the vector form of the equation (16.24):

$$f_1 = q\vec{\upsilon}\times\vec{B}. \tag{16.25}$$

As it can be seen from (16.25), this force is always perpendicular to the plane in which vectors $\vec{\upsilon}$ and $\vec{B}$ lie.

From mechanics it is known that if a force is perpendicular to velocit it changes only the direction of the latter, but not its value. Therefore, the Lorent force does not change the kinetic energy of a moving charge and does not do wor

If a charge is stationary relative to a magnetic field or its velocity is parall (antiparallel) to the magnetic induction vector, the Lorentz force is zero. Its direction [see (16.25)] depends on the sign of the charge.

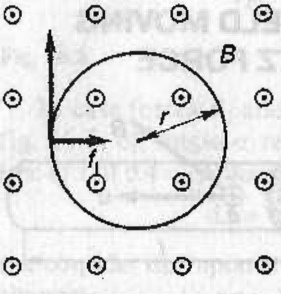


Fig. 16.10

Let a positively charged particle fly into a homogeneous magnetic field of induction B with velocity υ (fig. 16.10). It is under the action of Lorentz force f_l, which will cause centripetal acceleration, and, according to Newton's second law,

$$m\upsilon^2/r = q\upsilon B, \tag{16.26}$$

where q and m are the charge and mass; r is the radius of the trajectory along which it will be moving. From (16.26), we obtain:

$$r = m\upsilon/(qB). \tag{16.27}$$

From this, it follows that the radius of trajectory remains constant and the trajectory itself is a circumference.

Using (16.27) and assuming that the value of particle velocity does not change, we find the period of rotation along the circumference:

$$T = \frac{2\pi r}{\upsilon} = \frac{2\pi}{(q/m)B}. \tag{16.28}$$

The ratio q/m is called ***specific charge of particle***. The period of its rotatio in a magnetic field [see (16.28)] does not depend on the radius of the circum ference or velocity, but is determined only by magnetic induction and specif charge. This feature is used in the charged particle accelerator – *cyclotro* To describe the shape of the trajectory of a charged particle flying with velocity $\vec{\upsilon}$ into a homogeneous magnetic field at an arbitrary angle to $\vec{B}$ (fig. 16.11), let us decompose vector $\vec{\upsilon}$ into two components υ_{II} and $\upsilon_{\perp}$ directed along the magnetic induction vector of the magnet field and perpendicularly to it respective During the particle movement in the magnetic field, component υ_{II} remains constan the Lorentz force acting on the particle will change direction $\upsilon_{\perp}$.

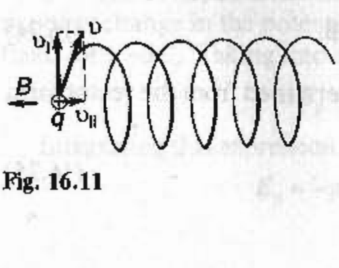


Fig. 16.11

Under the action of this force, the particle rotates along the circumference. Thus, the trajectory of motion will be a helix: rotation along the circumference together with motion along the vector of magnetic induction with velocity $\upsilon_{\parallel}$.

If onto a moving charged particle q is exposed to the effect of electric field with intensity E and magnetic field with magnetic induction B (fig. 16.12), then the resulting force is equal to:

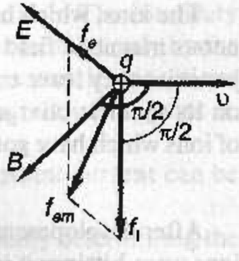


Fig. 16.12

$$\vec{f}_{em} = \vec{f}_e + \vec{f}_l = q\vec{E} + q\vec{\upsilon} \times \vec{B}. \tag{16.29}$$

In many systems (oscillograph, TV-set, electronic microscope), electrons or other charged particles are controlled by electric and magnetic fields. In this case, the main calculating formula is (16.29).

§ 16.4. EXPERIMENTAL DETERMINATION OF SPECIFIC CHARGE

Measurement of specific charge makes it possible to determine the mass of atoms and molecules and isotopic composition. Let us consider the operation principle of a device (fig. 16.13) used for this purpose. The flow of ions of the same sign flies through the electric and magnetic fields (the magnetic induction vector is directed everywhere from the reader perpendicularly to the plane of the drawing). Values E and B are chosen so that the fields act on charge q with forces equal in their absolute value but oppositely directed: $f_e = f_l$ or $qE = q\upsilon B$, whence,

$$\upsilon = E/B. \tag{16.30}$$

Some ions whose velocities satisfy condition (16.30) are not deflected by the fields and fly out of hole O, other ions are deflected (dashed lines in the figure) and caught. Thus, a part of the device shown in fig. 16.13 is the velocity selector; by changing E or B, from an ion beam, one can select groups with velocities determined by condition (16.30). A certain velocity spread is determined by the width of hole O.

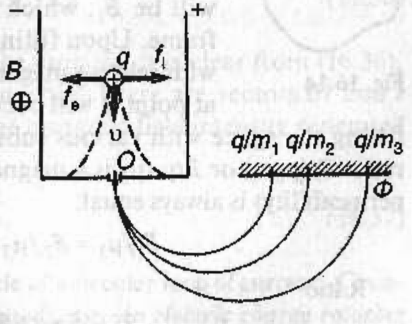


Fig. 16.13

The ions, which have flown from the velocity selector, get into a homogeneous magnetic field with induction B. Flying along half-circumference trajectories they leave traces on photo plate Φ in different parts of it depending on its specific charge. Using formula (16.27) we calculate specific charges of ions which have got into different parts of the photo plate:

$$q/m = \upsilon/(rB). \tag{16.31}$$

After development, the plate will have dark lines or spots in places where ions were hitting; it is possible to establish, first, the fact of presence of ions with a definite specific charge q/m or mass m, and, second, by intensity of the lines, to determine the part of ions with this or that value of their specific charge.

The device considered is a modification of mass spectrograph. In some cases separated ions record the flow; this version of the device is called ***mass-spectrometer.*** Mass spectrographs and mass-spectrometers are used for determining the isotope composition of a material.

§ 16.5. MAGNETIC FIELD INTENSITY. BIOT–SAVART–LA PLACE LAW AND ITS APPLICATION

In connection with the concept of magnetic induction, it is necessary to quantify this characteristic of the magnetic field depending on the actual configuration and value of currents in a medium. This problem leads to a certain auxiliary physical concept, which is ***magnetic field intensity.***

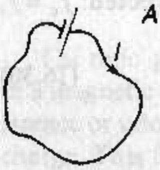


Fig. 16.14

Let us assume that at point A, there is a magnetic field generated by a loop of current with intensity I (fig. 16.14). If the entire space is filled with a material with relative magnetic permeability μ_{r1} then, at point A, magnetic induction will be B_1, which can be measured with a test magnetic frame. Upon filling the entire space with another material with relative magnetic permeability μ_{r2}, magnetic induction at point A will become B_2. Continuing this experiment and filling the space with various substances, it is possible to make sure that ratio $B/(\mu_0\mu_r)$ or B/μ (μ_0 is a magnetic constant, and μ is absolute magnetic permeability) is always equal:

$$B_1/\mu_1 = B_2/\mu_2 = \ldots = B/\mu. \tag{16.32}$$

Ratio

$$\vec{H} = \vec{B}/\mu = \vec{B}/(\mu_r\mu_0) \tag{16.33}$$

is called magnetic field *intensity* in a homogeneous medium. The intensuty does not depend on properties of the medium, but is determined only by the intensity of the current flowing along the loop, and geometric conditions of the experiment — shape of the loop and its position relative to point A. Vectors $\vec{H}$ and $\vec{B}$ coincide in their direction.

The intensity of a magnetic field generated by a constant current can be calculated using Biot–Savart–Laplace law.

J.B. Biot and F. Savart discovered this law experimentally determining the effect of currents of different shape on a magnetic needle. P.S. Laplace analyzed the data obtained by Biot and Savart and established that the magnetic field intensity of any current is the sum of the intensities of fields generated by its separate elements.

Let us take a conductor with current I and select current element $I\mathrm{d}\vec{l}$, from which we draw radius-vector $\vec{r}$ to point A (fig. 16.15; α is the angle between vectors $\mathrm{d}\vec{l}$ and $\vec{r}$). At point A, the element of current generates a magnetic field, elementary intensity $\mathrm{d}H$ of which is determined by the ***Biot–Savart–Laplace law***:

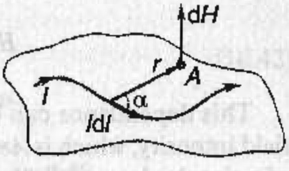


Fig. 16.15

$$\mathrm{d}H = k\frac{I\mathrm{d}l\sin\alpha}{r^2}. \tag{16.34}$$

In this, k is the proportionality coefficient depending on a choice of units. In the SI system $k = 1/(4\pi)$, hence

$$\mathrm{d}H = k\frac{I\mathrm{d}l\sin\alpha}{4\pi r^2}, \tag{16.35}$$

or in the vector form:

$$\mathrm{d}\vec{H} = k\frac{I}{4\pi r^3}\mathrm{d}\vec{l}\times\vec{r}, \tag{16.36}$$

According to the general rule of vector product, it is clear from (16.36), that $\mathrm{d}\vec{H}$ is perpendicular to the plane in which there are vectors $\mathrm{d}\vec{l}$ and r (fig. 16.15). Integrating (16.36), we find the magnetic field intensity generated by the loop of current or by its section:

$$\vec{H} = \frac{I}{4\pi}\int\frac{\mathrm{d}\vec{l}\times\vec{r}}{r^3}. \tag{16.37}$$

Magnetic field is in the center of the circle of a circular loop of current. *Circular current is one flowing along a circular conductor. An electric charge rotating along a circumference corresponds to this current.*

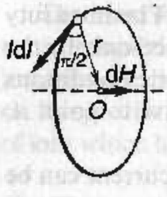


Fig. 16.16

In the circular current loop let us select current element Idl and set the direction of elementary intensity dH of the magnetic field generated by the element at point O which is the center of the circle of radius r (fig. 16.16). For any current element, dH is directed along the dotted line, which is perpendicular to the plane of the loop, therefore (16.37) can be written in the scalar form:

$$H = \frac{I}{4\pi}\int \frac{\sin\alpha\, dl}{r^2}.$$

Taking into account that $\sin\alpha = 1$ and $r = \text{const}$, we obtain an expression for the magnetic field intensity at the center of a circular loop:

$$H = \frac{I}{4\pi r^2}\int_0^{2\pi r} dl = \frac{I}{2r}. \tag{16.38}$$

This dependence can be used, in particular, to define a unit of magnetic field intensity, which is *Ampere per meter* (A/m). It is the intensity at the center of a circular loop with diameter 1 m if the current of 1 A flows along the loop.

Knowing the magnetic field intensity and relative magnetic permeability of a medium, we obtain magnetic induction:

$$B = \mu_r\mu_0 H = \mu_r\mu_0 I/(2r). \tag{16.39}$$

Magnetic field of an infinite rectilinear current conductor. As a certain abstraction, let us consider an infinite rectilinear current conductor (fig. 16.17, a). Let us select current intensity element Idl and draw radius-vector $\vec{r}$ to point A at distance b from the conductor. Vector dH is directed perpendicularly to the plane of the figure from the reader. The figure shows that relative to any current element, the elementary magnetic field intensity at point A is perpendicular to the plane of the drawing, so, like in the previous example (16.37), it can be written in the scalar form:

$$H = \frac{I}{4\pi}\int_{-\infty}^{+\infty} \frac{\sin\alpha\, dl}{r^2}, \tag{16.40}$$

Let us transform the integrand expression so that it contains only one variable — angle α.

From fig. 16.17, we obtain

$$r = b/\cos\beta \text{ and } \sin\alpha = \cos\beta. \tag{16.41}$$

In a larger scale, fig. 16.17 shows the angle, at which from point A, vector dl is visible. From triangle ACD, we can see that $|CD| - |CA|\,d\beta$. Since approxi-

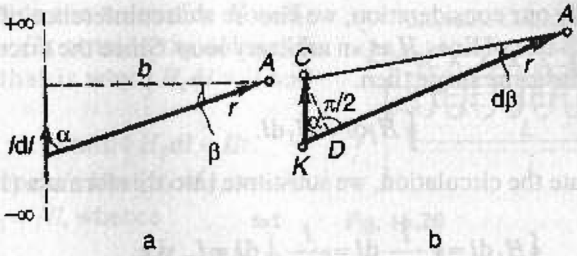


Fig. 16.17

mately $|CA| = r$, then $|CD| = r\mathrm{d}\beta$. From ΔCDK, we obtain $|CD| = \mathrm{d}l\sin\alpha$; therefore, $r\mathrm{d}\beta = \mathrm{d}l/\sin\alpha$, whence $\mathrm{d}l = r\mathrm{d}\beta/\sin\alpha$. Substituting (16.41) into the last equation, we obtain

$$Dl = b\,\mathrm{d}\beta/\cos^2\beta. \tag{16.42}$$

Taking into account (16.42), we transform (16.40):

$$H = \frac{I}{4\pi}\int\frac{\cos\beta\cos^2\beta\, b\,\mathrm{d}\beta}{b^2\cos^2\beta} = \frac{I}{4\pi b}\int_{-\pi/2}^{\pi/2}\cos\beta\,\mathrm{d}\beta.$$

Integration limits correspond to the extreme values of the angle: ±π/2. Integrating, we obtain an expression for a magnetic field intensity generated by an infinite rectilinear current conductor with current at any point at distance b from the conductor:

$$H = I/(2\pi b). \tag{16.43}$$

§ 16.6. AMPERE'S LAW. MAGNETIC FIELD INTENSITY OF A SOLENOID

In some cases, to determine the magnetic field intensity, it is expedient to use Ampere's law along with the associated Biot–Savart–Laplace law.

Let us draw lines of the intensity of a magnetic field[1] generated by an infinite rectilinear current conductor, which flows in a plane that is perpendicular to the conductor. The lines have a shape of concentric circumferences (fig. 16.18).

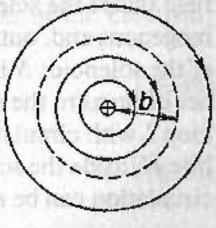


Fig. 16.18

---

[1] Tangents to these lines coincide with vectors $\vec{H}$.

To simplify our consideration, we choose a circumference of radius b that coincides with one of lines H as an arbitrary loop. Since the circuit and intensity line has the same shape then

$$\oint H_l \,\mathrm{d}l = H_l \,\mathrm{d}l. \tag{16.44}$$

To calculate the circulation, we substitute into this formula (16.43):

$$\oint H_l \,\mathrm{d}l = \oint \frac{I}{2\pi b}\,\mathrm{d}l = \frac{I}{2\pi b}\int_0^{2\pi b} \mathrm{d}l = I, \quad \text{i.e.}$$

$$\oint H_l \,\mathrm{d}l = \oint H \,\mathrm{d}l. \tag{16.45}$$

This is ***Ampere's law***, which connects the circulation of the vector of magnetic field intensity and current intensity.

A more complicated calculation could show that formula (16.45) is valid for an arbitrary circuit with any currents. Usually, this law is written down in the following form:

$$\oint H_l \,\mathrm{d}l = \sum I_i. \tag{16.46}$$

The circulation of the magnetic field intensity vector along the loop is equal to the algebraic sum of the currents flowing in this loop.

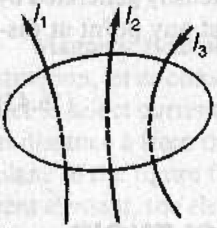


Fig. 16.19

Thus, if there are three currents in the loop (fig. 16.19) (currents *1* and *2* are positive, current *3* is negative), then Ampere's law for this case looks as

$$\oint H_l \,\mathrm{d}l = I_1 + I_2 - I_3.$$

Let us apply Ampere's law and find the magnetic field **intensity *of a solenoid*** (fig. 16.20; dotted lines indicate intensity). The longer a solenoid is and the smaller its diameter is, the more homogeneous the magnetic field inside the solenoid is. We assume that the field inside the solenoid is homogenous and, outside the solenoid, the field is quite weak. Let l be the length of the solenoid; N is the total number of turns; $n = N/l$ is the ratio of the number of turns to the solenoid length (winding density). Let us draw an arbitrary loop L with circulation given in fig. 16.20. One of its parts is l, it coincides with line H inside the solenoid, and another part l_1 goes beyond the solenoid. Thus, circulation can be expressed as a sum of two integrals:

$$\oint H_l \,\mathrm{d}l = \int_l H_l \,\mathrm{d}l + \int_{l_1} H_l \,\mathrm{d}l. \tag{16.47}$$

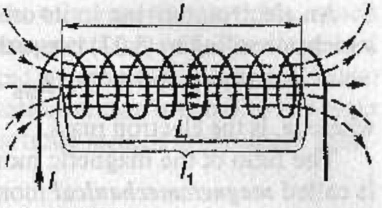


Fig. 16.20

Since, inside the solenoid, $H = H_1$, then $\int H_l \, dl = H_1$; outside the solenoid, H is small, that is why $\int\limits_{l_1} H_l \, dl = 0$ and from (16.47), we obtain $\oint H_l \, dl = Hl$.

Using Ampere's law (16.46), we obtain $H_1 = \sum I_1 = Nl$, whence

$$H = I(N/l) = In. \tag{16.48}$$

It means that the magnetic field intensity of a solenoid is equal to the product of current intensity by the number of turns divided by the length of the solenoid.

§ 16.7. MAGNETIC PROPERTIES OF A MATERIAL

There are no materials whose state would not change when they are placed into a magnetic field. Moreover, being in a magnetic field, materials themselves become sources of such a field. In this sense, all materials are called *magnetic materials*.

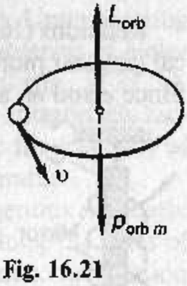


Fig. 16.21

Since macroscopic differences of magnetic materials are caused by their structure, it is expedient to discuss magnetic properties of electrons, nuclei, atoms and molecules as well as behavior of these particles in a magnetic field. We will treat the subject in the framework of classical physics.

Conventionally, we assume that in an atom, an electron moves around the nucleus uniformly with velocity υ along a circular orbit with radius r (fig. 16.21). This motion is similar to circular current and is characterized by orbital magnetic moment p_{orb} (one should keep in mind that an electron is a negatively charged particle and its motion is opposite to the current direction).

The intensity of the current corresponding to motion of the electron, which moves with frequency ν, is equal to

$$I = e\nu, \tag{16.49}$$

where e is the charge of the electron. Since $\nu = \upsilon/(2\pi r)$, then

$$I = e\upsilon/(2\pi r). \tag{16.50}$$

Then [see (16.2)]

$$p_{orb} = \frac{e\upsilon}{2\pi r}\pi r^2 = \frac{e\upsilon r}{2}. \tag{16.51}$$

An electron moving in its orbit has angular momentum L_{orb} (fig. 16.21), which according to (5.27) is equal to

$$L_{\text{orb}} = m_e \upsilon r, \tag{16.52}$$

where m_e is the electron mass.

The ratio of the magnetic moment of a particle to its angular momentum is called ***magnetomechanical*** moment. Dividing (16.51) by (16.52), we obtain ***orbital magnetomechanical ratio for an electron***:

$$G_{\text{orb}} = \frac{p_{\text{orb}}}{L_{\text{orb}}} = \frac{e}{2m_e}. \tag{16.53}$$

The magnetomechanical ratio is written down using the ***Lande' g-factor***:

$$G_{\text{orb}} = g_{\text{orb}} \frac{e}{2m_e} (g_{\text{orb}} = 1). \tag{16.54}$$

An electron also has its own angular momentum, which is called a *spin*. A spin corresponds to spin magnetic moment. ***Spin magnetomechanical*** ratio is twice higher than the orbital ratio:

$$G_s = \frac{e}{m_e} = g_s \frac{e}{2m_e} (g_s = 2). \tag{16.54a}$$

Relations (16.54) and (16.54a) show that between magnetic and mechanical (angular momentum) moments, there is a quite definite rigid relationship. Since e and m_e are constants[1], this relationship manifests itself in magnetomechanical phenomena. A. Eistein and de Haas first observed such a phenomenon in 1915. Light rod C was hung on a thin thread in the solenoid (fig. 16.22). As a current was passing through the solenoid, a magnetic field was generated and the magnetic moments of electrons were ordered, which led to an ordered orientation of angular momentums. As a result, the whole rod received an angular momentum and was turning — it was visible by deflection of a light spot reflected in the mirror.

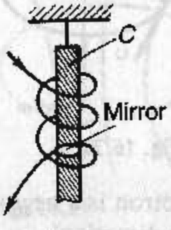


Fig. 16.22

Magnetomechanical phenomena make it possible to determine magnetomechanical relations and on the basis of this, to make conclusions of the importance of orbital or spin magnetic moments in processes of magnetization. Thus, the experiments of Einstein and de Haas showed that spin magnetic moments of electrons are responsible for magnetization of ferromagnetic materials.

Nuclei, atoms, and molecules also have a magnetic moment. The magnetic moment of a molecule is the vector sum of magnetic moments of atoms constituting it.

---

[1] The relativistic dependence of mass on velocity is not taken into account in this.

The magnetic field affects the orientation of a material particles, which have magnetic moments, so the material magnetizes itself. The degree of magnetization of a material is characterized by ***intensity of magnetization***. The average value of intensity of magnetization vector is equal to the ratio of the total magnetic moment $\sum \vec{p}_{mi}$ of all particles located in the space occupied by the magnetic to this volume:

$$\vec{J} = \frac{\sum \vec{p}_{mi}}{V}. \qquad (16.55)$$

Thus, intensity of magnetization is the average magnetic moment of the unit volume of the magnetic. The unit of the intensity of magnetization is *Ampere per meter* (A/m).

Magnetic materials are divided into three main classes: ***paramagnetic***, ***diamagnetic***, and ***ferromagnetic materials***. Each of them has its own type of magnetic properties: paramagnetism, diamagnetism, and ferromagnetism. Let us consider their nature.

According to the classical theory of paramagnetism, paramagnetic molecules have non-zero magnetic moments.

In absence of a magnetic field, these moments are located randomly, and the intensity of magnetization is equal to zero (fig. 16.23, a). Upon locating a paramagnetic sample into a magnetic field, the magnetic moments of molecules are mainly oriented in direction B, resulting in $J \neq 0$ (fig. 16.23, b). The degree of order of the magnetic moments depends on two antagonistic factors — the magnetic field and chaotic molecular motion, so the intensity of magnetization depends on both magnetic induction and temperature.

If a paramagnetic rod is hung in a vacuum in a homogenous magnetic field, then in the equilibrium position it will settle itself along the lines of magnetic induction (fig. 16.24, top view), which corresponds to orientation of $\vec{J}$ along the direction of $\vec{B}$. The magnetic field generated by paramagnetic material amplifies, although slightly, the external magnetic field; that is why induction $\boldsymbol{B}$ of the resulting field is higher than magnetic induction B_0

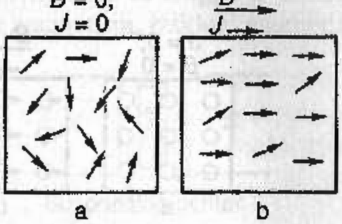


Fig. 16.23

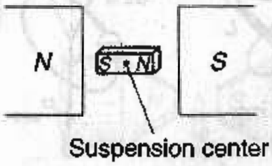


Fig. 16.24

of the field in absence of paramagnetic material ($B > B_0$). It means that relat magnetic permeability of paramagnetic materials is higher than 1 ($\mu_r>1$). Paramagnetic materials include aluminum, oxygen, molybdenum, etc. In a heterogonous magnetic field in a vacuum, particles of a paramagneti material move towards a higher value of magnetic induction, as they say, "a pulled into the field".

Explanation of the nature of diamagnetism is more complicated, so it is expedient to consider a mechanical phenomenon first.

No doubt, the reader observed that the axis of a common child's spinne makes cone-shaped rotational movements, called *precession* (fig. 16.25, a). It occurs when a rotating body with angular momentum L_{orb} is exposed to an destabilizing moment of force. If the spinner does not rotate, it overtur under the impact of the gravity force (mg) moment; rotation of the spinne generates precession.

A similar phenomenon takes place concerning electron orbits in a magne field. An electron moving in the orbit has an angular momentum like the spinner, and is also characterized by an orbital magnetic moment. That is wh a moment of force acts on it from the magnetic field like on a loop of curre Thus, conditions are created for occurrence of precession of the electron orbit or a rotating electron (fig. 16.25, b). Precession generates additional electr magnetic moment $\vec{P}_{\text{add}}$, directed opposite to induction $\vec{B}_0$ of the external magnetic field, and it weakens the total field. This is the nature of diamagnetis Diamagnetism is inherent to all materials. In paramagnetic materials, diamagnetism is repressed by stronger paramagnetism. If the magnetic moment of molecules is zero or is so small that diamagnetism predominates over paramagnetism; materials consisting of such molecules are diamagnetic materials.

Fig. 16.26 gives a schematic representation of diamagnetic molecules i absence of a magnetic field (a) and in its presence (b). The intensity of magnetization of diamagnetic materials is directed opposite to magnetic inductio its value increases if induction increases too.

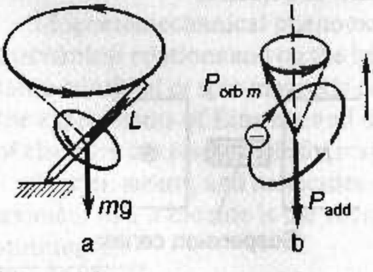


Fig. 16.25

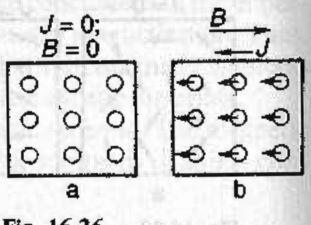


Fig. 16.26

Since self-magnetic field generated by a diamagnetic material is opposite to the external one, induction B inside such a material is less than induction B_0 in absence of the field ($B < B_0$). Therefore, relative magnetic permeability of a diamagnetic material is less than one ($\mu_r < 1$). Diamagnetic materials include nitrogen, hydrogen, copper, water, etc.

If a rod of diamagnetic material is suspended in a vacuum in a homogenous magnetic field, then in the equilibrium position, the rod will be set perpendicularly to the lines of magnetic induction (fig. 16.27, top view).

In a vacuum, in a heterogonous magnetic field, diamagnetic particles will be thrown out of the field. For instance, candle flame is deflected in such a field (fig. 16.28). Combustion products are diamagnetic particles.

Magnetic properties of materials depend on the structure of molecules, so magnetic measurement methods are used in chemical research. A specialized section of physical chemistry, magnetochemistry, studies relationships between magnetic and chemical properties of a material.

Ferromagnetic materials like paramagnetic ones generate magnetization directed along field induction; their relative magnetic permeability is much higher than one ($\mu_r \gg 1$). However ferromagnetism differs from paramagnetism considerably. Ferromagnetic properties are inherent not to individual atoms or molecules, but only to some materials in a crystalline state. Quantum theory explains this phenomenon.

Ferromagnetic materials include crystalline iron, nickel, cobalt, many alloys of these elements among themselves and with other non-ferromagnetic compounds, as well as alloys and compounds of chromium and manganese with non-ferromagnetic elements.

The intensity of magnetization of ferromagnetic materials depends not only on magnetic induction, but also on their previous state, the time of the sample's presence in a magnetic field. Ferromagnetic properties of a material are retained only below a certain temperature corresponding to the Curie point.

Although, there are not so many ferromagnetic materials in nature, they are mostly used as magnetic materials in technology. This is due to their strong magnetism, residual magnetization and coercive force.

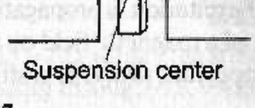


Fig. 16.27

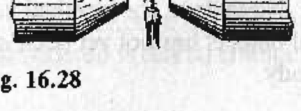


Fig. 16.28

Sizeable mechanical forces acting on ferromagnetic solids and permanent magnets in a magnetic field have diverse applications in medicine: correction of the thorax in children (Yu.F. Isakov, E.A. Stepanov, et al.), magnetic stoppers for prevention of secretions from an artificial stoma of colon (V.D. Fyodorov, T.S. Odaryuk, et al.), removing ferromagnetic particles and shavings from the eye.

§ 16.8. MAGNETIC PROPERTIES OF BODY TISSUES. PHYSICAL BASIS OF MAGNETOBIOLOGY

Body tissues are mainly diamagnetic like water. However, there are paramagnetic substances, molecules and ions in the human body.

Biocurrents that occur in a body are a source of weak magnetic fields. In some cases, induction of such fields can be measured. Thus, **based on registration of the time dependence of cardiac magnetic field induction (heart biocurrents), a diagnostic method** called ***magnetocardiography*** was developed.

Since magnetic induction is proportional to current intensity, and according to Ohm's law, biocurrent is proportional to voltage (bio-potential), a magnetocardiogram is generally similar to an electrocardiogram. However, magnetocardiography, in contrast to electrocardiography, is a non-invasive method, because a magnetic field can be registered at some distance from a biological object, which is the source of the field. Development of magnetocardiography depends on technical capabilities of measuring quite weak magnetic fields (see § 20.1).

A magnetic field affects biological systems that are located in it. This effect is studied by a section of biophysics called *magnetobiology*.

There are data that drosophilae die in a heterogeneous magnetic field, that morphological changes in animals and plants occur after exposure to a constant magnetic field, that plants acquire a certain orientation a magnetic field, and that the magnetic field affects the nervous system and changes in blood properties, etc.

Naturally, physical or physicochemical processes are primary in all cases.

Such processes can be orientating of molecules, a change in molecule or ion concentration in a heterogonous magnetic field, force action (the Lorentz force) on ions moving together with a biological liquid, the Hall effect occurring in a magnetic field when an electric pulse of excitation is propagating, etc.

Nowadays, the physical nature of the impact of a magnetic field on biological objects has not yet been established. This important problem is still under study.

Chapter 17

Electromagnetic induction. Energy of magnetic field

In the previous chapter, we showed an aspect of relationship between electric and magnetic phenomena: a moving electric charge is a source of a magnetic field. This chapter focuses on another aspect of this relationship: an alternating magnetic field generates an electric field; this is the essence of the phenomenon of electromagnetic induction, which was discovered by M. Faraday in 1831 after years of persistent research.

§ 17.1. BASIC LAW OF ELECTROMAGNETIC INDUCTION

Any change in the magnetic flux passing through a loop, generates an emf of electromagnetic induction in it.

Let a loop of a rectangular shape, located in the magnetic field of induction B, have a moving part of l long (fig. 17.1). When the current of intensity I flows along the loop, Ampere force F acts on the moving part and the latter moves to distance dx within a time interval dt. The work done by the current source is equal to the energy of heating according to the Joule–Lenz law and of displacement of the movable part:

$$\mathcal{E}I\mathrm{d}r = I^2R\mathrm{d}t + F\mathrm{d}x, \tag{17.1}$$

where $\mathcal{E}$ is emf of the source generating the current of intensity I in the loop; R is its electric resistance.

Since, in this example, the direction of the current is perpendicular to B (the cross in the circle in fig. 17.1 means that B is directed away from us beyond the drawing plane) then $F = IBl$. Substituting this expression into (17.1), we obtain $\mathcal{E}I\mathrm{d}t = I^2R\mathrm{d}t + IBl \cdot \mathrm{d}x$; d$t$ is assumed to be so small that current intensity can be considered to be constant. Cancelling by I, we obtain

$$\mathcal{E}\mathrm{d}t = IR\mathrm{d}t + Bl\mathrm{d}x. \tag{17.2}$$

Let us do some little transformations:

$$Bl\mathrm{d}x = B\mathrm{d}S = \mathrm{d}\Phi, \tag{17.3}$$

where $\mathrm{d}S = l\mathrm{d}x$ is the area shaded in fig. 17.1; dΦ is the change in the magnetic flux passing through dS, which in the example considered is due to change in the dimensions of the loop.

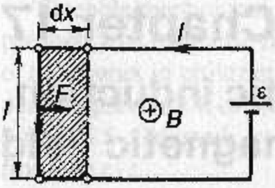


Fig. 17.1

Substituting (17.3) into (17.2), we obtain $\mathcal{E}\,dt = IR \cdot dt + d\Phi$, from where

$$I = \frac{\mathcal{E} - d\Phi/dt}{R}. \tag{17.4}$$

Equation (17.4) is ***Ohm's law for the whole loop.*** In the numerator of the right-hand member of the equation, is an algebraic sum of the electromotive forces in the loop: emf E of the source and emf $\mathcal{E}_i$ of electromagnetic induction, due to change in the magnetic flux passing through the loop. Thus,

$$\mathcal{E}_i = -d\Phi/dt. \tag{17.5}$$

It is the basic law of electromagnetic induction, or Faraday's law.

Although formula (17.5) is derived in a particular example, it is universal: when a magnetic flux passing through a loop changes (changes of the magnetic field in the course of time, advancing or removing the magnet, changes in the current in the adjacent loop or in the same one, etc.), this always generates emf of electromagnetic induction in the loop, which is proportional to the rate of magnetic flux change.

Expression (17.5) does not depend on any properties of the material from which the real loop is made; therefore, the emf of electromagnetic induction does not depend on these properties[1]. Moreover, we can even talk about electromotive force that is generated in an imaginary mathematical loop when the magnetic flux passing through it changes. This makes it possible to assume that a change in the magnetic field generates an electric field. The current, flowing in this case in a real conducting loop, is a result of action of an electric field.

Using Ohm's law for the whole loop and the basic law of electromagnetic induction, we will write the expression for induction current:

$$I = \frac{\mathcal{E}_i}{R} = -\frac{1}{R}\frac{d\Phi}{dt}. \tag{17.6}$$

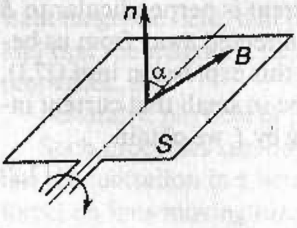


Fig. 17.2

From (17.6), it we can see that current intensity depends on the resistance of the loop; in a loop of ideal insulator, there will be no current.

If in a homogeneous magnetic field with induction B, a frame with area S rotates with angular velocity ω (fig. 17.2) uniformly, then on the basis of (16.7) and (5.5), we can write down the law of change in the magnetic flux passing through the frame:

---

[1] The material, from which the real loop is made, can affect the emf if its magnetic permeability depends on induction of themagnetic field. This is not taken into account here.

$$\Phi = BS\cdot\cos\alpha = BS\cdot\cos\omega t.$$

From (17.6) for this case we have:

$$I = -\frac{1}{R}\frac{\mathrm{d}}{\mathrm{d}t}(BS\cdot\cos\omega t) = I_{\max}\sin\omega t. \qquad (17.7)$$

$I_{\max} = BS\omega/R$ is the current intensity amplitude. As we can see from (17.7), in this case, alternating sinusoidal induction current is generated in the frame.

The minus sign in formula (17.6) reflects the *Lenz rule*.

Let us explain it with the following example. Let the magnet draw near the loop by the north pole (fig. 17.3). Let us suppose that the direction of bypass of the loop shown in the figure (small arrows) is positive, then normal n to the area element, encircled with the loop, has the corresponding direction (assuming the system is the right-handed). From the figure, we can see that angle α is obtuse and, therefore [see (16.7)], the magnetic flux is negative[1]. When the magnet is approaching the loop, magnetic induction B increases, the absolute value of the magnetic flux passing through the loop increases, but since the flux is negative, then $\Delta\Phi < 0$ and $\mathrm{d}\Phi/\mathrm{d}t < 0$. By formula (17.6), we obtain that $I > 0$. It means that the direction of induction current coincides with the direction of bypass along the loop (large arrows in fig. 17.3), which does correspond to the Lenz rule. Let us determine the charge flowing in the conductor due to electromagnetic induction. Since the current intensity is a derivative of the charge by time, then, taking in account (17.6), we can write:

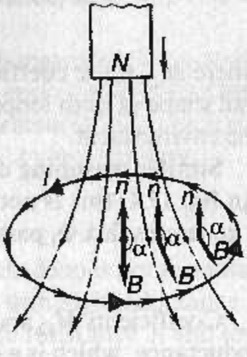


Fig. 17.3

$$\frac{\mathrm{d}q}{\mathrm{d}t} = -\frac{1}{R}\frac{\mathrm{d}\Phi}{\mathrm{d}t}, \quad \text{or} \quad \mathrm{d}q = -\mathrm{d}\Phi/R. \qquad (17.8)$$

Integrating the last equation, we obtain:

$$\int_0^q \mathrm{d}q = -\frac{1}{R}\int_{\Phi_1}^{\Phi_2} \mathrm{d}\Phi, \quad q = \frac{\Phi_2 - \Phi_1}{R} = -\frac{\Delta\Phi}{R}. \qquad (17.9)$$

From this it follows that a charge flowing in the conductor due to electromagnetic induction, depends on changes in the magnetic flux passing through the loop and on its resistance.

---

[1] In this, formula (16.7) can be applied here conventionally since the magnetic field is heterogeneous.

In particular, this dependence is used to measure magnetic flux with instruments recording the electrical charge stimulated in the loop.

§ 17.2. MUTUAL INDUCTION

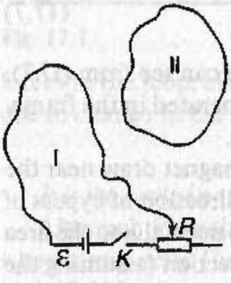


Fig. 17.4

Generation of emf in one loop when the current flowing through the other loop changes is called mutual induction. This is a particular case of electromagnetic induction.

Let us consider two loops I and II (fig. 17.4). If we close switch K, then in loop I, current I_1 will occur generating a magnetic field. Magnetic flux Φ_2 will pass through loop II. It is proportional to magnetic induction, and it is proportional to the current generating the magnetic field.

Therefore,

$$\Phi_2 = M_{21} I_1, \qquad (17.10)$$

where M_{21} is the coefficient of proportionality depending on the dimensions and shape of both loops, their mutual location and magnetic permeability of the environment.

Similar reasoning can be made upon reversing the roles of loops I and II (in fig. 17.4, this is not show). If current of intensity I_2 flows along loop II, it generates flux Φ_1 passing through loop I:

$$\Phi_1 = M_{12} I_2.$$

Coefficients M_{12} and M_{21} are equal: $M_{12} = M_{21} = M$; they are called mutual inductance, which is a characteristic of two circuits (loops). The unit of their mutual inductance is henry (H). From the definition it is clear that 1 H = = 1 W/1 A.

When the current intensity in the first loop changes, the magnetic flux passing through the second one changes too, generating an emf of electromagnetic induction. Using the basic law of electromagnetic induction and ratio (17.10), we write down, considering M to be a constant:

$$\mathcal{E}_{i2} = -\frac{d\Phi_2}{dt} = -M\frac{dI_1}{dt}. \qquad (17.11)$$

Similarly, for the emf generated in the first circuit, we obtain

$$\mathcal{E}_{i1} = -\frac{d\Phi_1}{dt} = -M\frac{dI_2}{dt}. \qquad (17.12)$$

As we can see from (17.11) and (17.12), the emf of mutual induction depends on the rate of change in current intensity in the adjacent loop and mutual inductance of both loops.

§ 17.3. SELF-INDUCTION

In a circuit, generation of an emf of induction due to a change in current intensity of the same circuit is called self-induction. It is also a particular case of electromagnetic induction.

We can conclude that in this case the relationship between the intensity of current flowing along the loop and the magnetic flux generated by it and passing through the loop is similar to (17.10):

$$\Phi = LI, \tag{17.13}$$

where L is the circuit (loop) inductance.

Inductance depends on the dimensions and shape of the loop and magnetic permeability of the medium. The unit of inductance, as well as that of mutual inductance, is henry.

When the current intensity in the loop changes, the magnetic flux passing through this loop changes too, which generates an emf of self-induction. Using the basic law of electromagnetic induction and formula (17.13), we obtain an expression for an emf of self-induction, considering L to be a constant:

$$\mathcal{E}_i = -\frac{\mathrm{d}(LI)}{\mathrm{d}t} = -L\frac{\mathrm{d}I}{\mathrm{d}t}. \tag{17.14}$$

Due to a self-induction, a change in the current in the loop is not instantaneous. Let us show it with an example (fig. 17.5). A coil with high inductance L and low resistance together with the resistor with high resistance R and low inductance are connected in parallel to the source. When switch K is disconnected, the upper part of the circuit (a) will become an independent loop (b), in which the current is generated only by the electromotive force of self-induction.

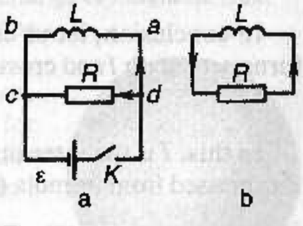


Fig. 17.5

Based on Ohm's law for the whole loop and formula (17.14)

$$IR = -L\frac{\mathrm{d}I}{\mathrm{d}t} \quad \text{or} \quad \frac{\mathrm{d}I}{I} = -\frac{R}{L}\mathrm{d}t. \tag{17.15}$$

Let us integrate the last equation, assuming that to the initial moment of time $t = 0$, current intensity I_0 corresponds, and to the intermediate moment

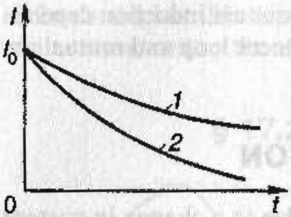


Fig. 17.6

of time t, current intensity I corresponds:

$$\int_{I_0}^{I}\frac{dI}{I}=-\frac{R}{L}\int_0^t dt;\quad \ln\frac{I}{I_0}=-\frac{R}{L}t. \tag{17.16}$$

Finding the antilogarithm we obtain:

$$I=I_0\exp[-(R/L)t]. \tag{17.17}$$

Therefore, in the loop (fig. 17.5, b), the current intensity does not change instantly, but does so exponentially. That is why, in particular, upon disconnecting any real loop, there is a spark or arc on the terminals of the switch.

In fig. 17.6, there are time dependencies on the current intensity for two values R/L[1]:

$$(R/L)_1<(R/L)_2.$$

Such processes are called transient ones; they occur in electrical circuits when passing from one mode of operation to another.

Theoretically, such a process, as it follows from (17.17), is infinite. However, it is accepted to estimate the duration by time τ, during which the parameter characterizing a transient process (in the example considered, it is the current intensity) changes e times (the time constant).

For the time-constant, the expression can be derived from (17.17), if instead of I, we substitute I_0/e and instead of t we substitute τ: $I_0/e = I_0\exp[-(R/L)\tau]$, from where

$$\tau=L/R. \tag{17.17a}$$

For transient processes (fig. 17.6),

$$\tau_1=\tau_2.$$

In conclusion, let us calculate the inductance of the solenoid having N turns per length l and cross-section area S. From (17.3), we have

$$L=\Phi/I. \tag{17.18}$$

In this, I is the intensity of the current flowing along the solenoid, which is expressed from formula (16.48):

$$I=Hl/N. \tag{17.19}$$

---

[1] Formula (17.17) and curves (fig. 17.6) are valid under the assumption that the initial current intensity at segment ab is much higher than the initial current intensity in segment dc (fig. 17.5, a). When we disconnect switch K, the direction of the current and its intensity i in segment dc change.

The magnetic flux[1], passing through all its N loops is:

$$\Phi = BSN = \mu_{r0} HSN. \tag{17.20}$$

Substituting (17.19) and (17.20) into (17.18), we obtain:

$$L = \frac{\mu_r \mu_0 HSN^2}{Hl} = \mu_r \mu_0 \frac{N^2 S}{l}. \tag{17.21}$$

From this, we can see that L depends on geometrical properties of the loop (N, S, l) and its relative magnetic permeability μ_r of the medium inside the solenoid.

Multiplying the numerator and denominator in (17.21) by length l of the solenoid and taking into account that $Sl = V$ (solenoid volume) and $n = N/l$, we obtain:

$$L = \mu_r \mu_0 n^2 V. \tag{17.22}$$

From formula (17.22), it is easy to determine that the unit of μ_0 is a ***henry per meter*** (H/m).

§ 17.4. EDDY CURRENTS

One of manifestations of electromagnetic induction is generation of closed induction currents (eddy currents, or Foucault's currents) in continuous conductive bodies: metal parts, electrolyte solutions, biological organs, etc.

Eddy currents occur when a conductive body moves in a magnetic field, when the induction of the field changes over time, as well as under the cumulative impact of both factors. As we can see from (17.6), the intensity of eddy currents depends on the electrical resistance of a body and, therefore, on specific resistance and dimensions, as well as on the rate of change of magnetic flux.

Let us cite several examples of some applications of eddy currents.

Eddy currents, according to the Joule–Lenz law, cause heating of conductors, which are used for casting metals in special furnaces and heating the surface of conducting bodies with the purpose of their surface quenching.

In physiotherapy, heating individual human body parts with eddy currents is administered as a treatment called inductothermy 17.7 (see § 19.3).

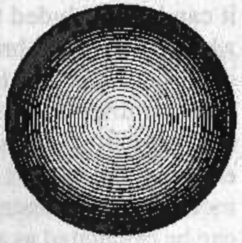


Fig. 17.7

---

[1] In electrical engineering, a flow passing through all or several turns of a solenoid is called a flux linkage.

Upon interaction of eddy currents generated in moving conductors with the magnetic field, conductors are decelerated. It is used, for example, in needle electric measurement instruments for deceleration of their moving parts to fasten reading.

Nevertheless, in certain cases, the action of eddy currents is undesirable. For instance, heating the cores of transformers, motors, and other devices is associated with nonproductive energy consumption, and it is necessary sometimes to cool these components. To reduce undesirable heating, the electric resistance of the cores is intentionally increased by using silicon steel plates or ferrite materials.

Eddy currents manifested themselves in a peculiar way in filaments of alternating current circuits. In this case, AC intensity is redistributed about the conductor cross-section: the current intensity is lower in internal parts of the cross-section and higher in its external parts. Fig. 17.7 sketchily shows distribution of the alternating current intensity about the cross-section of a cylindrical conductor: the thickness of the rings is approximately proportional to the current density at different distances from its axis. This phenomenon is known as ***skin effect***[1]. The degree of current displacement to the conductor surface depends on the current frequency. Thus, for high frequency currents, almost the entire inner part of the conductors has no current, so it is expedient to use hollow conductors (tubes), and, in some cases, to cover them with a good but expensive conductor, such as silver.

§ 17.5. ENERGY OF MAGNETIC FIELD

As it was already noted in § 17.3, when the circuit is disconnected, a spark springs up; this indicates energy conversion, because the spark is a source of radiation, heating, and sonic waves. Since in this case the current intensity decreases and, therefore, the induction of the magnetic field decreases too, it can be concluded that the current and induction generate energy, which, when the circuit is broken, converts to other forms of energy. To calculate this energy, let us return to the circuit in fig. 17.5.

After disconnection of switch K in the loop consisting of L and R, the current generated by the emf of self-induction, flows, and the magnetic field energy of the current is converted mostly into the energy of molecular thermal motion, which heats the resistor. The decrease in the magnetic field energy can be calculated as a work of this current:

$$-\Delta E_m = A_I.$$

---

[1] Skin (in this context) means a surface effect.

The final values of the current intensity, magnetic field induction and energy are equal to zero; the initial value of the magnetic field energy of the current will be E_m, that is why $\Delta E_m = 0 - E_m = -E_m$ and then:

$$E_m = A_i. \tag{17.23}$$

The elementary work done by the current during time interval dt is equal to $dA_i = \mathcal{E}_i I \cdot dt$. Substituting (17.14) in the last expression, we have:

$$dA = -L\frac{dI}{dt}I \cdot dt = -L \cdot I \cdot dI.$$

Integrating this equation within the limits of the current intensity variation from I to 0, we obtain:

$$E_m = \int dA_i = -L\int_I^0 I \cdot dI = \frac{LI^2}{2}. \tag{17.24}$$

Formula (17.24) is valid for any circuit; it shows the dependence of the magnetic field energy of the current on its intensity in the circuit (loop) and on its inductance.

It is useful to compare formula (17.24) with the expression for kinetic energy of translational motion: $m\upsilon^2/2$. This analogy shows that the inductance of a loop characterizes its certain inertial properties. Just as it is impossible to stop a moving body without energy conversions so it is impossible to stop the motion of charges (of the current) without energy conversion.

Let us express the energy of a magnetic field in terms of its characteristics with an example of a solenoid. Let us assume that the magnetic field of a long solenoid is homogeneous and located mostly inside it. Substituting I from (17.19) and L from (17.22) into (17.24), we obtain:

$$E_m = \frac{\mu_r \mu_0 N^2 V H^2 l^2}{l^2 2N^2} = \frac{\mu_r \mu_0 H^2}{2}V. \tag{17.25}$$

Dividing both parts of (17.25) by the solenoid volume V, we obtain the volume energy density of the magnetic field:

$$w_m = \mu_r \mu_0 H^2/2 \tag{17.26}$$

or taking into account (16.33):

$$w_m = BH/2. \tag{17.27}$$

The unit of volumetric energy density is a *Joule per cubic meter* (J/m^3).

Expression (17.26) is valid for any magnetic field irrespective of its origin and determines magnetic field energy related to volume.

It is useful to compare (17.26) with formula (14.46) for the volumetric density of the energy of electric field. These expressions are similar, though $\overline{E}$ and $\overline{H}$ are not the same.

Chapter 18

Electromagnetic oscillations and waves

Electromagnetic oscillations are periodic (or almost periodic) interrelated changes in charge, current intensity, intensity of electric and magnetic fields. In space, propagation of electromagnetic oscillations occurs in the form of electromagnetic waves. Among various physical phenomena, electromagnetic oscillations and waves occupy a special place. Almost all electrical engineering, radio-engineering as well as optics are based on these concepts.

§ 18.1. FREE ELECTROMAGNETIC OSCILLATIONS

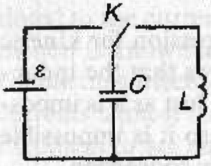


Fig. 18.1

Free (proper) electromagnetic oscillations are those that occur without any external impact due to the energy stored initially.

Let us consider a closed RLC circuit consisting of inductor L and capacitor C (fig. 18.1), which is charged from source $\mathcal{E}$ by switch K, and then is discharged into the inductance coil. At all that, in the circuit, an emf of self-induction occurs which will be equal to the voltage on the capacitor plates. Using formula (17.14), we write down:

$$-L(dI/dt)=q/C. \qquad (18.1)$$

From this, taking into account (15.3), we obtain:

$$-L\frac{d^2q}{dt^2}=\frac{q}{C} \quad \text{or} \quad \frac{d^2q}{dt^2}+\omega_0^2 q=0, \qquad (18.2)$$

where

$$\omega_0^2=1/(LC). \qquad (18.3)$$

It is known that (18.2) is a differential equation of a harmonic oscillation; its solution [see (7.6)] has the form:

$$q=q_{\max}\cos(\omega_0 t+\varphi_0), \qquad (18.4)$$

where $q_{\max}$ is the maximum (initial) charge on the capacitor plates; ω_0 is the circular frequency of proper oscillations (***natural circular frequency***) of the circuit; φ_0 is the initial phase.

Not only the charge on the plates of the capacitor changes according to the harmonic law, but also the voltage and current intensity in the circuit respectively:

$$U = \frac{q}{C} = \frac{q_{\max}}{C}\cos(\omega_0 t + \varphi_0) = U_{\max}\cos(\omega_0 t + \varphi_0), \tag{18.5}$$

$$I = \frac{dq}{dt} = q_{\max}\omega_0 \sin(\omega_0 t + \varphi_0) = -I_{\max}\sin(\omega_0 t + \varphi_0), \tag{18.6}$$

where $U_{\max}$ and $I_{\max}$ are amplitudes of the voltage and current intensity.

Curves of the dependence of the charge and voltage on time are similar to the curve of $x = f(t)$, and the curve of the dependence of current intensity on the time is similar to the curve of function $\upsilon = f(t)$ (see fig. 7.4).

From (18.3), we find an expression for the period of proper oscillations (***Thomson's formula***):

$$T = \frac{2\pi}{\omega_0} = 2\pi\sqrt{LC}. \tag{18.7}$$

The energy of the charged capacitor [see (14.43)] and the energy of the magnetic field of a circuit [see (17.24)] change periodically:

$$E_{el} = (CU_{\max}^2/2)\cos^2(\omega_0 t + \varphi_0), \tag{18.8}$$

$$E_{m} = (LI_{\max}^2/2)\sin^2(\omega_0 t + \varphi_0). \tag{18.9}$$

It is possible to make certain that the sum of expressions (18.8) and (18.9) does not depend on time: **in an ideal circuit, the total energy is conserved, electric oscillations are undamped.**

A real RLC circuit (fig. 18.2) has ohmic resistance, so in it, the oscillations attenuate. For this circuit, Ohm's law can be written as:

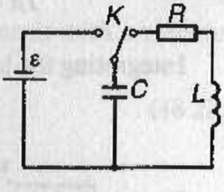


Fig. 18.2

$$-L(dI/dt) = IR + q/C. \tag{18.10}$$

Transforming this equation:

$$\frac{d^2q}{dt^2} + \frac{R}{L}\frac{dq}{dt} + \frac{1}{LC}q = 0 \tag{18.11}$$

and making a substitution:

$$2\beta = R/L, \quad \omega_0^2 = 1/(LC), \tag{18.12}$$

we obtain:

$$\frac{d^2q}{dt^2} + 2\beta\frac{dq}{dt} + \omega_0^2 q = 0. \tag{18.13}$$

This equation coincides with the differential equation of damped mechanical oscillations [see (7.33)]. If their attenuation is not too high, i.e. $\omega_0 - \beta = = \omega > 0$, we find the following solution [see (7.34)]:

$$q = q_{\max} e^{-\beta t} \cos(\omega_0 t + \varphi_0). \tag{18.14}$$

The curve of this function is similar to the curve in fig. 7.14. If attenuation is low ($\omega_0^2 \gg \beta^2$), then $\omega \approx \omega_0$. In this case, the logarithmic decrement of attenuation

$$\lambda = \beta T = \frac{R}{2L} \cdot 2\pi\sqrt{LC} = \pi R \sqrt{\frac{C}{L}}. \tag{18.15}$$

$\omega_0^2 < \beta^2$, or using (18.12),

$$\frac{1}{LC} < \frac{R^2}{4L^2}. \tag{18.16}$$

Let us consider two particular cases following from inequation (18.16).

1. There is no capacitor in the circuit; in other words, the capacitive resistance of the circuit (see § 18.2) is zero. As we can see from (18.28), it means formally that $C \to \infty$. From (18.10), we obtain equation (17.15) whose solution is presented in the form of (17.17) and its curve is shown in fig. 17.6.

2. There is no inductor in the circuit ($L \to 0$). From (18.10), we have

$$IR = -\frac{q}{C}, \quad \frac{dq}{dt} = -\frac{q}{C}, \quad \text{or} \quad \frac{dq}{q} = \frac{dt}{RC}. \tag{18.17}$$

Integrating the latter equation, we obtain:

$$\int_{q_{\max}}^{q} \frac{dq}{dt} = -\int_{0}^{t} \frac{dt}{RC}, \quad \ln \frac{q}{q_{\max}} = \frac{t}{RC}. \tag{18.18}$$

Finding the antilogarithm of the second expression (18.18), we have:

$$q = q_{\max} \exp[-t/(RC)]. \tag{18.19}$$

Equation (18.19) describes the process of discharging the capacitor onto resistor R. On the capacitor plates, the voltage changes by meeting the same law. In the absence of inductance, oscillations do not occur (fig. 18.3, a).

We can show that from the source of constant emf $\mathcal{E}$, charging the capacitor meets the exponential law:

$$q = \mathcal{E}C\{1 - \exp[-t/(RC)]\} = q_{\max}\{1 - \exp[-t/(RC)]\}. \tag{18.20}$$

The curve for this equation is shown in fig. 18.3, b. For a circuit with a capacitor and resistor, the transient process time constant (see § 17.3) is:

$$\tau = RC. \tag{18.21}$$

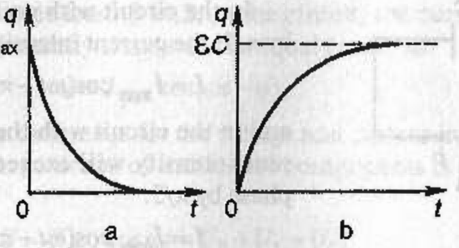


Fig. 18.3

Electromagnetic oscillations are often referred to as electrical ones.

§ 18.2. ALTERNATING CURRENT

In a broad sense, alternating current is any current that changes over time. However, the term of "alternating current" is more often applied to currents that change over time according to the harmonic law. An alternating current can be considered fo be forced electromagnetic oscillations.

Let us imagine three different circuits (fig. 18.4, a–18.6, a) to each of which an alternating voltage is applied:

$$U = U_{\max} \cos \omega t. \tag{18.22}$$

Let us note without proof that the current in the circuit with a resistor (fig. 18.4) will change in the phase with applied voltage:

$$I = I_{\max} \cos \omega t. \tag{18.23}$$

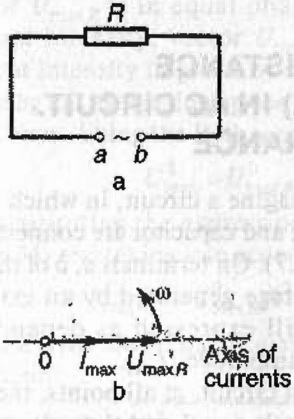


Fig. 18.4

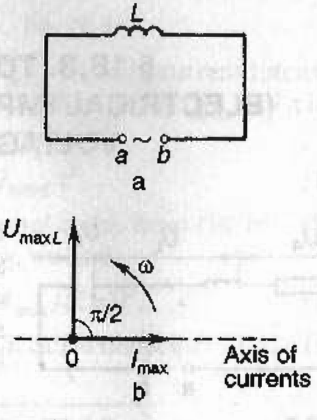


Fig. 18.5

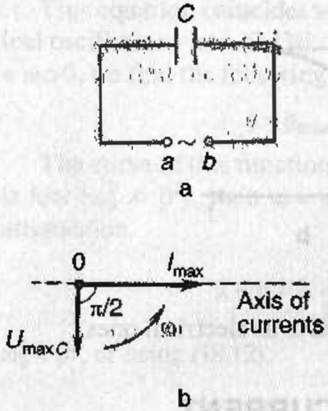


Fig. 18.6

In the circuit with coil inductance shifted by $\pi/2$, the current intensity is:

$$I = I_{\max}\cos(\omega t - \pi/2), \qquad (18.24)$$

and in the circuit with the capacitor, the current intensity will exceed the voltage in the phase by $\pi/2$:

$$I = I_{\max}\cos(\omega t + \pi/2). \qquad (18.25)$$

The vector diagrams corresponding to these examples are shown in fig. 18.4, b–18.6, b.

According to Ohm's law, the ratio of voltage amplitude ($U_{\max}$) to the current intensity amplitude ($I_{\max}$) expresses resistance.

For a circuit with a resistor, we have ***ohmic resistance***:

$$R = \frac{U_{\max}}{I_{\max R}}, \qquad (18.26)$$

for a circuit with inductor we have ***inductive resistance***:

$$X_L = \frac{U_{\max}}{I_{\max L}} = L\omega, \qquad (18.27)$$

for a circuit with a capacitor we have ***capacitive resistance***:

$$X_C = \frac{U_{\max}}{I_{\max C}} = \frac{1}{C\omega}. \qquad (18.28)$$

§ 18.3. TOTAL RESISTANCE (ELECTRICAL IMPEDANCE) IN AC CIRCUIT. VOLTAGE RESONANCE

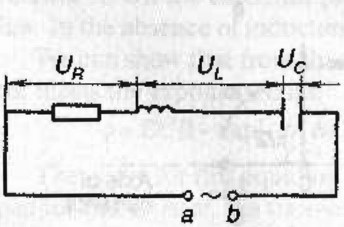


Fig. 18.7

Let us imagine a circuit, in which resistor, inductor, and capacitor are connected in series (fig. 18.7). On terminals *a, b* of the circuit, the voltage generated by an external source is still expressed as dependence (18.22) with amplitude $U_{\max}$.

In a serial circuit, at all points, the current intensity is equal and the voltages are different. As we can see from § 14.2, in the circuit, the current intensity and the voltage do not generally change in the equal phase, so

$$I = I_{\max}\cos(\omega t - \varphi), \tag{18.29}$$

where φ is the phase difference for the voltage and current intensity.

The sum of voltage drops of individual components is equal to external voltage:

$$U = U_{\max}\cos\omega t = U_R + U_L + U_C. \tag{18.30}$$

In accordance with § 18.2, U_R, U_L, U_C can be written as:

$$U_R = U_{\max R}\cos(\omega t - \varphi) \text{ (in the phase with current intensity);} \tag{18.31}$$

$$U_L = U_{\max L}\cos(\omega t - \varphi + \pi/2) \tag{18.32}$$
(exceeds current intensity by the phase);

$$U_C = U_{\max C}\cos(\omega t - \varphi - \pi/2) \tag{18.33}$$
(lags behind the current intensity in phase).

Substituting (18.31)–(18.33) into (18.30), after trigonometric transformations, it is possible to obtain an expression for the impedance of the AC circuit and angle φ. However, it can be done more simply and clearly using vector diagrams.

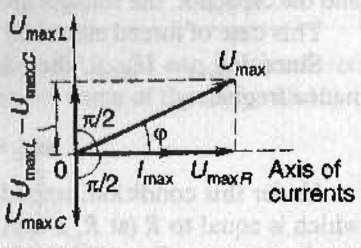


Fig. 18.8

In fig. 18.8, the vector of the current intensity amplitude $I_{\max}$ is directed along the axis of currents. Since the amplitude of the current intensity is equal along the whole circuit, we draw the individual voltage drop amplitudes relative to this vector: vector $U_{\max R}$ is in equal phase with the current intensity; vector $U_{\max L}$ exceeds current intensity in phase by $\pi/2$, vector $U_{\max C}$ exceeds current intensity in phase by $\pi/2$. Summing up the three vectors, we find values $U_{\max}$ and π in diagram form. Using the Pythagorean theorem, we have

$$U_{\max}^2 = U_{\max R}^2 + (U_{\max L} - U_{\max C})^2. \tag{18.34}$$

Substituting the expressions of these amplitudes from (18.26)–(18.28) in (18.34) and taking into account Ohm's law, we find:

$$I_{\max}^2 Z^2 = I_{\max R}^2 R^2 + [I_{\max} L\omega - I_{\max}/(C\omega)]^2, \tag{18.35}$$

where Z is the total resistance of the AC circuit called impedance. From (18.35) we obtain:

$$Z = \sqrt{R^2 + [L\omega - 1/(C\omega)]^2} = \sqrt{R^2 + (X_L - X_C)^2}. \tag{18.36}$$

Ohmic resistance R of the circuit is also called active, it causes releas of heat in a circuit in accordance with the Joule–Lenz law. The differenc between inductive and capacitive resistance $(X_L - X_C)$ is called reactive resistance. It does not cause heating of the electric circuit elements.

Let us write down Ohm's law applied to the amplitudes of the voltage and current intensity in the circuit (fig. 18.7):

$$I_{\max} = U_{\max}/Z = U_{\max}/\sqrt{R^2 + (X_L - X_C)^2}. \tag{18.37}$$

From fig. 18.8, we will also find value φ expressed through the known values

$$\begin{aligned}\tan\varphi &= \frac{U_{\max L} - U_{\max C}}{U_{\max R}} = \frac{I_{\max} L_{\omega} - I_{\max}/(C\omega)}{I_{\max R}} = \\ &= \frac{L_{\omega} - 1/(C\omega)}{R} = \frac{X_L - X_C}{R}.\end{aligned} \tag{18.38}$$

If inductive and capacitive resistance of the circuit are equal $(X_L = X_C)$ upon a serial connection, then [see (18.36)] $Z = R$. From (18.13) we have $\tan\varphi = 0$ and $\varphi = 0$. It means that the current intensity and the voltage applied change in the same phase, as if there was only Ohmic resistance in the circuit; on the inducto and the capacitor, the voltages are equal in their amplitude but opposite in phase

This case of forced electrical oscillations is called voltage ***resonance.***

Since $U_{\max L} = U_{\max C}$, then $L\omega_{\text{res}} = 1/(C\omega_{\text{res}})$. From this, we find the resonance frequency:

$$\omega_{\text{res}} = \sqrt{1/LC}. \tag{18.39}$$

Under this condition, impedance Z of the circuit has the minimum value which is equal to R (at R, L and C given), and the current intensity attains its maximum value. For voltage resonance in a circuit, the vector diagram is shown in fig. 18.9. If $L\omega > 1/(C\omega)$, then $\tan\varphi > 0$ and $\varphi > 0$, the current intensity lags behind the phase of applied voltage (see fig. 18.8). If $L\omega < 1/(C\omega)$, then we have $\tan\varphi < 0$ and $\varphi < 0$. The current intensity exceeds the voltage in phase.

For this case, a vector diagram is shown in fig. 18.10.

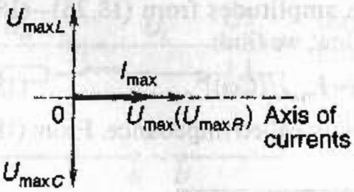


Fig. 18.9

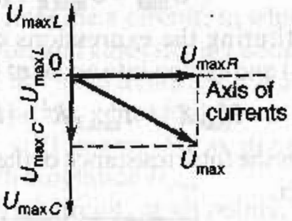


Fig. 18.10

§ 18.4. TOTAL RESISTANCE (IMPEDANCE) OF BODY TISSUES. PHYSICAL PRINCIPLES OF RHEOGRAPHY

Body tissues conduct not only direct current (DC), but also alternating current (AC). There are no systems in the body similar to inductors, so its inductance is close to zero. Biological membranes and, therefore, the whole body have capacitive properties; in this connection, only ohmic and capacitive resistance determine the impedance of body tissues. In biological systems, the presence of capacitive elements is proved by the fact that the current intensity exceeds the applied voltage in phase. We give some values of the phase shift angle obtained at frequency 1 kHz for different biological objects (Table 18.1).

Table 18.1

Object	φ, degree
Human, frog skin	55
Frog nerve	64
Rabbit muscles	65

Ohmic and capacitive properties of biological tissues can be simulated using equivalent electrical circuits. Let us consider some of them (fig. 18.11).

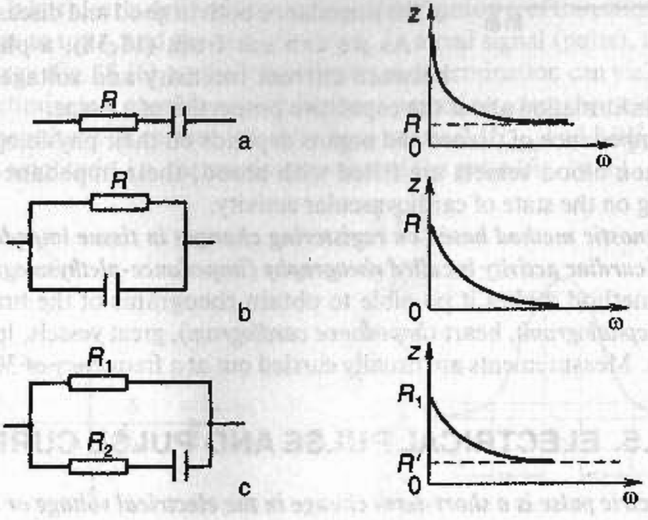


Fig. 18.11

For the diagram shown in fig. 18.11, a, the impedance frequency dependence can be obtained from (18.36) when $L = 0$:

$$Z = \sqrt{R^2 + 1/(C\omega)^2}. \qquad (18.40)$$

The curve shows that the diagram has an essential deviation from the experiment data at $\omega \to 0$ ($Z \to \infty$). Does it turn out that the resistance is infinitely high when we have direct current? This contradicts the data given in § 15.4.

The equivalent circuit diagram (fig. 18.11, b) differs from the experimental data when $\omega \to \infty$. In fact, at high frequencies, biological tissues still have resistance.

The most successful circuit diagram is an equivalent one combining the first two models (fig. 18.11, c). When $\omega \to \infty$ $X_C \to 0$, the resistance of the circuit can be found using the rule of parallel connection of resistances:

$$R = R_1R_2/(R_1 + R_2).$$

A dependence of the impedance on the frequency make it possible to assess viability of organ tissues, which is important to know for transplantation of tissues and organs. Let us illustrate it graphically (fig. 18.12). In this, *1* is the curve for good, normal tissue; *2* is for dead tissue killed by boiling water. In the dead tissue membranes — "live capacitors" —, are destroyed and the tissue has only ohmic resistance.

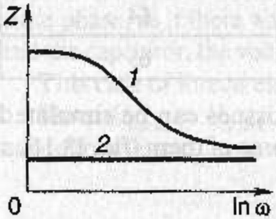


Fig. 18.12

There is difference in the frequency dependence of the impedance both in good and diseased tissues.

As we can see from (18.38), a phase angle between current intensity and voltage drop can also give information about the capacitive properties of a tissue.

The impedance of tissues and organs depends on their physiological state. Thus, when blood vessels are filled with blood, their impedance changes depending on the state of cardiovascular activity.

A diagnostic method based on registering changes in tissue impedance in the process of cardiac activity is called rheography (impedance-plethysmography).

This method makes it possible to obtain rheograms of the brain (*impedance encephalogram*), heart (*impedance cardiogram*), great vessels, lungs, liver, and limbs. Measurements are usually carried out at a frequency of 30 kHz.

§ 18.5. ELECTRICAL PULSE AND PULSE CURRENT

An electric pulse is a short-term change in the electrical voltage or current.

In technology, pulses are divided into two large groups: video- and radio pulses:

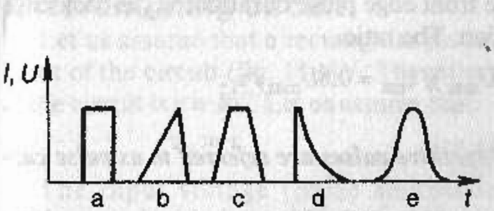


Fig. 18.13

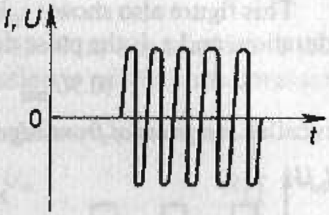


Fig. 18.14

Video pulses are such electrical pulses of current or voltage that have a constant component, which is not equal to zero. Thus, a videopulse has mainly one polarity. In their shape, the video pulses are (fig. 18.13):

a) rectangular;
b) saw-tooth;
c) trapezoidal;
d) exponential;
e) bell-shaped, etc.

Radio pulses are electromagnetic modulated oscillations (fig. 18.14).

In physiology, the term "electric pulse" or "electric signal" refers exactly to video pulses, so let us consider parameters of these pulses estimating their shape, duration, and the properties of their individual segments.

Characteristic segments of a pulse (fig. 18.15) are: *1–2* is the front edge, *2–3* is the top, *3–4* is the back edge, *4–5* is the tail. The pulse in this figure is quite schematic. It has clearly defined time points of beginning t_1 of transition from the leading edge to top t_2 and the end of pulse t_5. In a real signal (pulse), these time data are fuzzy (fig. 18.16), so their experimental determination can yield sizeable errors. To eliminate a possible error it has been agreed to select the points of time at which the voltage or current intensity have values $0.1U_{max}$ and $0.9U_{max}$, where U_{max} is the amplitude, i.e. the maximum value of the pulse (fig. 18.16).

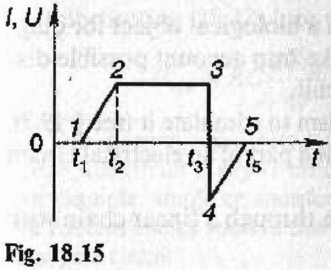


Fig. 18.15

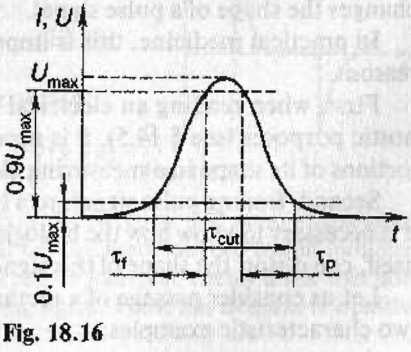


Fig. 18.16

This figure also shows: τ_f is the front edge pulse duration; τ_{cut} is back edg duration, and τ_p is the pulse duration. The ratio

$$(0.9U_{max} - 0.1U_{max})/\tau_{cut} = 0.8U_{max}/\tau_f.$$

is called ***steepness of front edge.***

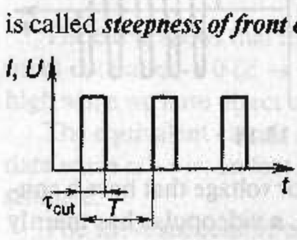


Fig. 18.17

Repetitive pulses are referred to as pulse current.

It is characterized by period (pulse repetitio period) T, which is the average time between the beginnings of adjacent pulses (fig. 18.17) and frequency (pulse repetition frequency) $f = 1/T$. *Period-to-pulse duration ratio* is:

$$Q = T/\tau_p = 1/f\tau_p. \quad (18.41)$$

The inverse ratio between Q is a ***duty cycle***:

$$K = 1/Q = f\tau_p. \quad (18.42)$$

§ 18.6. PASSAGE OF RECTANGULAR PULSES THROUGH A LINEAR CIRCUIT. DIFFERENTIATING AND INTEGRATING CIRCUITS

When an alternating current passes through an electrical circuit consisting of resistors, inductors and capacitors, the signal retains its harmonic shape a sinusoidal electric current corresponds to this external harmonic signal. Thus, there is a linear relationship between the current intensity and voltage and the circuit itself is called *linear*. In the circuit, the presence of such elements as the tube, semiconductor diode, transistor, makes the circuit *nonlinear*.

The linear circuit does not distort the shape of harmonic voltage wave, but changes the shape of a pulse signal.

In practical medicine, this is important to take into account for two main reasons.

First, when reading an electrical signal from a biological object for diagnostic purposes (see § 14.5), it is necessary to take into account possible distortions of its shape in a measuring electrical circuit.

Second, feeding pulse signals to a living organism to stimulate it (see § 19.2), it is necessary to know how the biological system, as a part of an electrical circuit itself, can distort the shape of this signal.

Let us consider passage of a rectangular pulse through a linear chain with two characteristic examples.

Differentiating circuit

Let us assume that a rectangular pulse (fig. 18.19, a) has been fed to the input of the circuit (fig. 18.18)[1]. The pulse duration is τp, the time constant of the circuit is $\tau = RC$. Let us assume that

$$\tau_p \gg \tau. \tag{18.43}$$

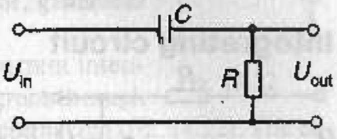


Fig. 18.18

The input voltage (pulse amplitude) is distributed between the capacitor (U_C) and resistor (U_R):

$$U_{in} = U_{max} = U_C + U_R. \tag{18.44}$$

The capacitor is charged when a pulse is fed [see (18.19)] and discharged after its termination [see (18.20)] meeting the exponential law. An approximate time dependence $U_C(t)$ is shown with a curve (fig. 18.19, b). The output voltage is equal to the voltage drop across the resistor:

$$U_{out} = U_R = IR. \tag{18.44a}$$

From (18.44), it follows:

$$U_{out} = U_R = U_{in} - U_C. \tag{18.45}$$

The curve of time dependence $U_R(J)$ can be plotted by subtracting curves b from a (see fig. 18.19, a).

Condition (18.43) makes it possible to conclude that shortly after the occurrence of the pulse in the circuit $U_{out} \ll U_C$ (fig. 18.19), in this case from (18.45) we have:

$$U_{in} \approx U_C. \tag{18.46}$$

Since $I = \frac{dq}{dt}$, $q = U_C C$, then $I = C\frac{dU_C}{dt}$, and taking into account (18.46):

$$I \approx C\frac{dU_{in}}{dt}. \tag{18.47}$$

Substituting (18.47) into (18.44 a), we obtain:

$$U_{out} \approx RC\frac{dU_{in}}{dt}. \tag{18.48}$$

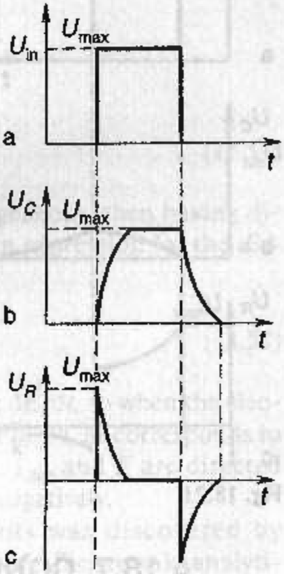


Fig. 18.19

---

[1] This diagram represents a four-pole circuit. In electrical and radio engineering, a *four-pole* circuit is any electrical circuit that has two input and two output terminals (for example, amplifier, transformer, etc.). A *passive* four-pole circuit is one that does not contain energy sources that affect a passing signal. Thus, this diagram is a passive four-pole circuit.

As we can see from (18.48), with the assumptions made, in this diagra the output voltage is proportional to a time derivative of the input voltag So, this circuit is called a ***differentiating*** one.

Integrating circuit

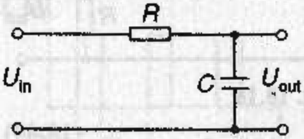


Fig. 18.20

We have another linear circuit (fig. 18.20) whose input parameter is a rectangular pulse (fig. 18.21, a). Let the pulse duration be much less than the time constant of the circuit:

$$\tau \gg \tau_p. \qquad (18.49)$$

In this case, the capacitor is charged more slowly (fig. 18.21, b). Like in the previous example, $U_R(t)$ can be found by subtracting one curve from the other (fig. 18.21, c), since formula (18.44) is valid here, too.

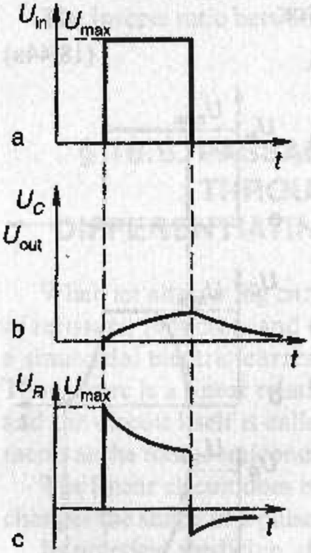


Fig. 18.21

The output voltage is the voltage on the capacitor:

$$U_{in} = U_C.$$

From condition (18.49) it follows that $U_{in} \gg U_C$, we can also see this from fig. 18.21, a, b. In this case, from (18.44), we obtain the approximate equality $U_{in} \approx U_R$. Since

$$U_R = IR, \text{ то } U_{in} \approx RC\frac{dU_{out}}{dt}.$$

From this,

$$U_{out} \approx \frac{1}{RC}\int U_{in}\,dt. \qquad (18.50)$$

So, this circuit is called ***integrating*** circuit. In both examples, the shape of a pulse has been changed (distorted) very much.

§ 18.7. CONCEPT OF MAXWELL'S THEORY. DISPLACEMENT CURRENT

Summarizing the results of experiments of H.C. Ørsted on electric current impact onto a magnetic needle, Faraday's experiments on magnetic induction, and other facts, Maxwell developed the theory of an electromagnetic field inside the framework of classical physics.

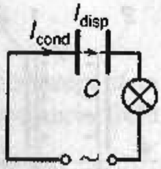


Fig. 18.22

Maxwell's theory is based on two following ideas.

1. Every alternating electric field generates an eddy magnetic field. Maxwell called an alternating electric field a displacement current since it, like a regular current, generates a magnetic field.

To find an expression for the displacement current intensity, let us consider passage of an alternating current through a circuit, to which a capacitor with a dielectric is connected (fig. 18.22). The capacitor does not impede flow of the current, which is indicated by the bulb. In conductors, this is a usual conduction current I_{cond} due to a change in the charge on the capacitor plates. We can assume that in the capacitor, the conduction current turns into displacement current with intensity I_{disp}, and

$$I_{disp} = I_{cond}\frac{dq}{dt}.$$

On capacitor plates, the charge is

$$q = CU = \frac{\varepsilon_r \varepsilon_0 S}{l} El = \varepsilon_r \varepsilon_0 SE.$$

Then the displacement current intensity is

$$I_{disp} = \varepsilon_r \varepsilon_0 \frac{dE}{dt} S. \qquad (18.51)$$

Since the electric field of the capacitor is homogeneous, then having divided the current intensity by the area we obtain an expression for the ***displacement current density***:

$$\vec{j}_{disp} = \varepsilon_r \varepsilon_0 \frac{d\vec{E}}{dt}. \qquad (18.52)$$

From (18.52), it follows that $\vec{j}_{disp}$ is directed along $d\vec{E}/dt$, so when the electric field intensity increases, it is directed along $\vec{E}$. Fig. 18.22 corresponds to this case: on capacitor plates, its charge increases, $\vec{j}_{disp}$ and $\vec{E}$ are directed from the plate charged positively to the one charged negatively.

The magnetic field of displacement currents was discovered by W.C. Roentgen experimentally and established by A.A. Eichenwald analytically[1].

In Eichenwald's experiment, a disk from a dielectric *1* (fig. 18.23) is located between the plates of two flat capacitors *2* and *3*. In them, the electric field

---

[1] A.A. Eichenwald was the first head of the physics' sub-faculty of the Higher Women's Courses in Moscow, based on which a number of Moscow universities were established, including the Russian Medical University.

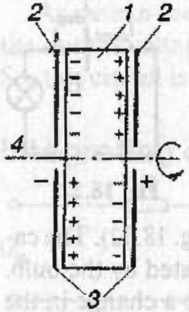


Fig. 18.23

intensities are antiparallel. Upon rotation of the disc around axis *4*, the polarization of the dielectric changes in the space between the capacitors. This phenomenon generates a magnetic field detected with a special indicating magnetic needle.

Substituting the expression for the displacement current intensity (18.51) into Ampere's law (16.46), we obtain the first ***Maxwell equation***:

$$\oint H_l \mathrm{d}l = I_{\text{cond}} + \varepsilon_r \varepsilon_0 \frac{\mathrm{d}E}{\mathrm{d}t} S, \qquad (18.53)$$

which shows that **an eddy magnetic field is generated both by conductive currents with intensity I_{cond} (moving electric charges) and displacement currents (alternating electric field B).**

2. Every alternating magnetic field generates an eddy electric field (*the basic law of electromagnetic induction*). Using expression (15.5) and law (17.5), we obtain ***the second Maxwell equation***:

$$\oint E_l \mathrm{d}l = -\frac{\mathrm{d}\Phi}{\mathrm{d}t}, \qquad (18.54)$$

which interrelates the rate of change in the magnetic flux through any surface and circulation of the electric field intensity vector generated in this process. The circulation is taken along the circuit, on which the surface rests.

From the main ideas of Maxwell's theory, which are given above, it follows that at a certain point in space, generation of any field, be it electric or magnetic, entails a chain of mutual transformations: an alternating electric field a magnetic field (fig. 18.24, a shows the lines of generated magnetic field generated if $\mathrm{d}E/\mathrm{d}t > 0$); a change in the magnetic field generates an electric field (fig. 18.24, b shows the fluxes of generated electric field, if $\mathrm{d}H/\mathrm{d}t > 0$), etc. The difference in the sign of the Maxwell equations (18.53) and (18.54) determines the different direction of the arrows on the lines of $\vec{H}$ and $\vec{E}$ in these figures.

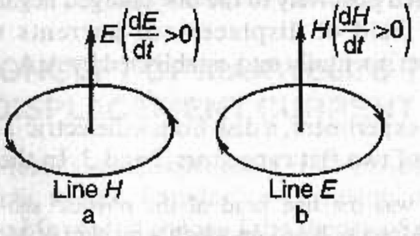

$E\left(\frac{dE}{dt}>0\right)$
$H\left(\frac{dH}{dt}>0\right)$
Line H
a
Line E
b

Fig. 18.24

§ 18.8. ELECTROMAGNETIC WAVES

Mutual formation of electric and magnetic fields leads to the concept of an ***electromagnetic wave***: propagation of the unified electromagnetic field In space.

Let us explain this. Let us assume that at point x_1 of a dielectric (fig. 18.25), the electric field intensity $\vec{E}_1$ increases. This generates an eddymagnetic field whose intensity $\vec{H}_2$ at point x is directed from the reader (compare with fig. 18.24, a). Increasing of $\vec{H}_2$ generates an eddy electric field; at point x_2, the intensity vector of this field is perpendicular to axis *OX* (compare with fig. 18.24, b), etc. If changes in $\vec{E}$ or $\vec{H}$ are maintained at this point at the expense of energy from a certain source, then an electromagnetic wave will propagate in space continuously.

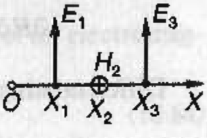


Fig. 18.25

Let us show that the wave nature of electromagnetic field propagation follows from Maxwell's equations (18.53) and (18.54). We will consider the medium to be a dielectric; therefore, the conductivity current intensity is zero. The magnetic flux through a certain area *S* perpendicular to $\vec{B}$ is equal to:

$$\Phi = BS = \mu_r\mu_0 SH.$$

Under these conditions, Maxwell's equations will be as follows[1]:

$$\oint H_l \mathrm{d}l = \varepsilon_r\varepsilon_0 S\frac{\partial E}{\partial t}; \quad \oint E_l \mathrm{d}l = -\mu_r\mu_0 S\frac{\partial H}{\partial t}. \tag{18.55}$$

On plane *ZOX*, let us select (fig. 18.26) an elementary, that is a quite small rectangular loop *1–2–3–4–1*. The area bounded by the loop is $\mathrm{d}x\,\mathrm{d}z$. On segments *1–2* and *3–4*, $El\,\mathrm{d}I = 0$; on segment *2–3* $El\,\mathrm{d}I = -E\partial z$, on segment *4–1* $El\mathrm{d}I = (E + \partial \mathrm{E})\partial z$. Hence, the circulation of vector $\vec{E}$ along this loop is:

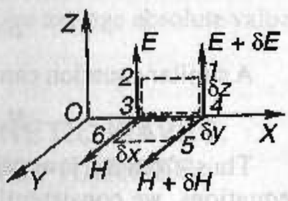


Fig. 18.26

$$0 + (-E\partial z) + 0 + (E + \partial E)\partial z = \partial E\,\partial z.$$

Substituting the expressions for the area and the circulation into the second equation of (18.55), we obtain:

$$\partial E\,\partial z = \mu_r\mu_0 \partial x \partial z\frac{\mathrm{d}H}{\mathrm{d}t} \quad \text{or} \quad \frac{\partial E}{\partial x} = -\mu_r\mu_0\frac{\partial H}{\partial t}. \tag{18.56}$$

---

[1] Maxwell's equations are represented in partial derivatives since later it will be necessary to differentiate by coordinates.

Similarly, considering rectangular loop *3–6–5–4–3* selected on plan *XOY*, you can notice that on segment *3–6*, $Hdl = H\partial y$, on segments *6–5* and *4–3*, $Hdl = 0$; on segment *5–4*, $Hdl = (H + \partial H)\partial y$. From this the circulatio of vector $\vec{H}$ along this segment is:

$$H\partial y + 0 + [-(H + \partial H)\partial y] = \partial H\,\partial y.$$

Substituting the expressions for the area bounded by the circuit and circulation into the first equation of (18.55), we obtain:

$$-\partial H\partial y = \varepsilon_r\varepsilon_0\partial x\partial z\frac{dE}{dt} \quad \text{or} \quad \frac{\partial H}{\partial x} = -\varepsilon_r\varepsilon_0\frac{\partial E}{\partial t}. \qquad (18.57)$$

Differentiating (18.56) by coordinate x and (18.57) by time, we obtain:

$$\frac{\partial^2 E}{\partial x^2} = \mu_r\mu_0\frac{\partial^2 H}{\partial x\partial t} \quad \text{or} \quad \frac{\partial^2 H}{\partial x\partial t} = \frac{1}{\mu_r\mu_0}\frac{\partial^2 E}{\partial x^2},$$

$$\frac{\partial^2 H}{\partial x\partial t} = -\varepsilon_r\varepsilon_0\frac{\partial^2 E}{\partial t^2}.$$

From the last equations, it follows:

$$\varepsilon_r\varepsilon_0\frac{\partial^2 E}{\partial t^2} = \frac{1}{\mu_r\mu_0}\frac{\partial^2 E}{\partial x^2} \quad \text{or} \quad \frac{\partial^2 E}{\partial x^2} = \varepsilon_r\varepsilon_0\mu_r\mu_0\frac{\partial^2 E}{\partial t^2}. \qquad (18.58)$$

This is ***a wave equation*** (7.49), let us represent its solution in the form of a flat wave equation (7.45):

$$E = E_m\cos\omega(t - x/\upsilon). \qquad (18.59)$$

A similar equation can be derived for the magnetic field intensity:

$$H = H_m\cos\omega(t - x/\upsilon). \qquad (18.60)$$

Thus, from the laws of Biot–Savart–Laplace's and Faraday, with Maxwe equations, we consistently come to the wave nature of the electromagne field propagation.

Comparing (18.58) with wave equation (7.49) we conclude that the velocit of the electromagnetic wave propagation is:

$$\upsilon = \frac{1}{\sqrt{\varepsilon_r\varepsilon_0\mu_r\mu_0}} = \frac{c}{\sqrt{\varepsilon_r\mu_r}}, \qquad (18.61)$$

where $c = 1/\sqrt{\varepsilon_0\mu_0}$ is the velocity of light in vacuum.

Thus, *the velocity of electromagnetic wave propagation equals the spee of light.* This served as the basis of *of Maxwell's electromagnetic theory of ligh*

Comparing (18.61) with the expression for a refraction index $n = c/\upsilon$ it is possible to determine the connection between n, relative dielectric and magnetic permeability:

$$n = \sqrt{\varepsilon_r \mu_r}. \qquad (18.62)$$

The volumetric energy density of an electromagnetic field consists of volume densities of the electric (14.46) and magnetic (17.26) fields:

$$\omega_{\text{em}} = \omega_{\text{el}} + \omega_{\text{m}} = \varepsilon_r \varepsilon_0 E^2/2 + \mu_r \mu_0 H^2/2. \qquad (18.63)$$

In a dielectric, the electric and magnetic components of an electromagnetic field are energetically equivalent, so:

$$\varepsilon_r \varepsilon_0 E^2/2 = \mu_r \mu_0 H^2/2, \qquad (18.64)$$

then for the volumetric energy density, we can write several expressions:

$$\omega_{\text{em}} = \varepsilon_r \varepsilon_0 E^2 = \mu_r \mu_0 H^2 = \sqrt{\varepsilon_r \varepsilon_0 \mu_r \mu_0}\, EH. \qquad (18.65)$$

The density of the wave energy flux (wave intensity) is obtained from the general formula, substituting (18.61) and (18.65) into it:

$$I^2 = \sqrt{\varepsilon_r \varepsilon_0 \mu_r \mu_0}\, EH \frac{1}{\sqrt{\varepsilon_r \varepsilon_0 \mu_r \mu_0}} = EH, \qquad (18.66)$$

or

$$\vec{I} = \vec{E} \times \vec{H}. \qquad (18.67)$$

In this formula, E and H are understood to be the time average absolute value of these vectors.

§ 18.9. SCALE OF ELECTROMAGNETIC WAVES. CLASSIFICATION OF FREQUENCY RANGES USED IN MEDICINE

From Maxwell's theory, it follows that various electromagnetic waves, including light waves, have a common nature. In this connection, it is expedient to represent all kinds of electromagnetic waves in a unified scale (fig. 18.27).

The whole scale is conventionally divided into six ranges: ***radio waves*** (long, medium, and short ones), ***infrared***, ***visible***, ***ultraviolet***, ***X-rays***, and ***gamma-radiation*** (γ). This classification is made according to the mechanism of wave formation or the possibility of visual perception of waves by people.

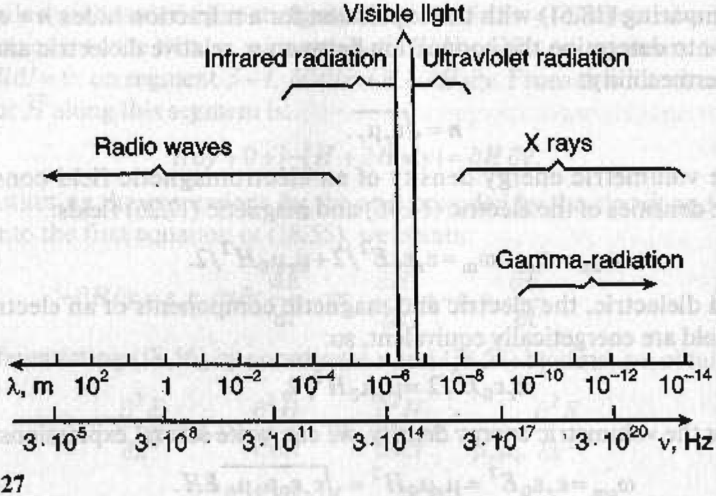


Fig. 18.27

Radio waves are due to alternating currents in conductors and electron flows (macro-emitters). Infrared, visible, and ultraviolet radiation occurs in atoms, molecules, and fast charged particles (micro-emitters). X-ray is generated by intra-atomic processes, γ-radiation is of nuclear origin.

Certain ranges overlap because waves of equal length can occur in different processes. Thus, the shortest wave ultraviolet radiation is overlapped by long-wave X-ray radiation.

In this respect, the boundary region of infrared waves and radio waves is quite characteristic. Before 1922, there was a gap between these ranges. The shortest-wave radiation of this gap was of molecular and atomic origin (radiation from a heated body); macroscopic Hertz's vibrators emitted the long-wave radiation. The Russian physicist A.A. Glagoleva-Arkadieva[1] suggested passing a spark through a mixture of a large number of metal filings in oil. Various electromagnetic waves with a wavelength of 82 μm or more could be obtained. The infrared and radio wave ranges were united.

Nowadays, no one is surprised that even millimeter waves can be generated not only by radio equipment, but also by molecular transitions. A section of physics, which studies absorption and emission of radio waves by various materials, has sprung up; its name is *radiospectroscopy.* In medicine, the following conventional division of electromagnetic oscillations into frequency ranges is accepted (Table 18.2).

---

[1] Alexandra Andreevna Glagoleva-Arkadieva was the first Head of the Sub-faculty of Physics at the 2nd Moscow Medical School (nowadays, it is the Russian State Medical University).

Table 18.2

Low (LF)	Up to 20 Hz
Audio (AF)	20 Hz–20 kHz
Ultrasonic (USF)	20 kHz–200 kHz
High (HF)	200 kHz–30 MHz
Ultrahigh (UHF)	30 MHz–300 MHz
Super-high (SHF)	300 MHz–300 GHz
Extremely high (EHF)	Over 300 GHz

Physiotherapeutic electronic equipment of low and audio frequency is often referred to as low frequency. Electronic equipment of all other frequencies is generally referred to as *high frequency.*

Chapter 19

Physical processes in tissues under action of current and electromagnetic fields

All materials consist of molecules; each of them is a system of charges, so the state of bodies essentially depends on the currents flowing through them and on the acting electromagnetic field. Electrical properties of biological bodies are more complicated than properties of inanimate objects because an organism is also a totality of ions with a variable concentration in space. The primary mechanism of the effect of currents and electromagnetic fields on a body is physical; in the chapter, it is considered as applied to therapeutic methods.

§ 19.1. PRIMARY ACTION OF DIRECT CURRENT (DC) ON BODY TISSUES. GALVANIZATION. ELECTROPHORESIS OF MEDICINAL SUBSTANCES

The human body consists largely of biological fluids containing a large number of ions that are involved in various metabolic processes.

Under the influence of an electric field, ions move at different velocities and accumulate near cell membranes, forming a counter electric field called *polarization field*. Thus, the primary effect of direct current is connected with motion of ions, their separation, and a change in their concentration in different components of tissues.

The effect of a direct current on the body depends on its intensity, so the electrical resistance of tissues and primarily the skin is very significant. Moisture and sweat reduce resistance very much, which even at a small voltage, can cause a sizeable current passing through the body.

Continuous DC with a voltage of 60–80 V is used as a physiotherapeutic method (*galvanization*).

The current source is usually a two half-wave rectifier — a device for galvanization. This requires electrodes of sheet lead or tinfoil 0.3–0.5 mm thick. Since products of electrolysis of a sodium chloride solution contained in the tissues cause burning pain, hydrophilic pads moistened, for example, with warm water are placed between the electrodes and the skin.

The current intensity is dosed based on milliammeter indications, always taking into account the maximum permissible current density of 0.1 mA/cm^2.

Direct current is also used in medical practice, for administration of medications through the skin or mucous membranes. This method is called *electrophoresis of medications.*

They do it similarly to galvanization, but the pad of the active electrode is moistened with a solution of the corresponding medication.

The medication is administered from the pole whose charge it has: anions are administered from the cathode; cations from the anode.

Administration of medicinal substances using DC is well illustrated by the following experiment. In two rabbits, spots of skin are shaved on both sides and flannel pads are attached to the shaved places; some pads are moistened with solution of strychnine nitrate, others are moistened with sodium chloride solution (fig. 19.1). Electrodes are applied to the flannel pads, and a current of 50 mA passes through the circuit. In some time, the rabbit with strychnine on the anode dies with typical symptoms of poisoning by this substance. The other rabbit, which has strychnine on the cathode, does not die, but if we change the direction of the current, then it will die too.

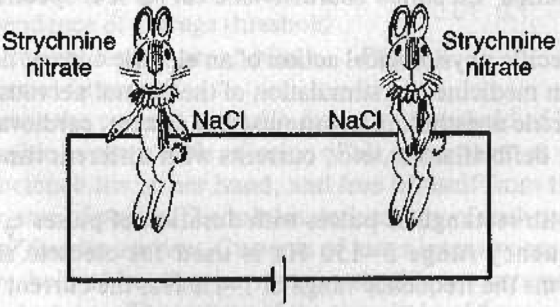


Fig. 19.1

Galvanization and electrophoresis of medications can be carried out using liquid electrodes in the form of baths, in which patient's extremities are immersed.

§ 19.2. EXPOSURE TO ALTERNATING (PULSE) CURRENTS

The effect of alternating current on the body depends on its frequency very much. At low, acoustic and ultrasonic frequencies (see § 18.9), alternating current as well as direct current causes an irritation of biological tissues. Thi is due to displacement of ions of electrolyte solutions, their separation, changes in their concentration in different parts of a cell and its intercellular spac

Tissue irritation depends on the shape of pulse current, duration of the pulse and its amplitude. For example, increasing the steepness of the puls front edge reduces the threshold current intensity which causes muscles contraction. This indicates that muscles adapt to the change in the current intensity, and ion compensation processes begin. The steepness of a rectangul pulse is very high (theoretically, it is infinit so for such pulses, the threshold current intensi is lower than for other ones. There is a definit relationship between threshold amplitude I_{ma} and rectangular duration of a pulse, which cause irritation (fig. 19.2). There is a corresponden between each point of the curve, points above the curve, and pulses that cause muscles contractio Points located below the curve correspond to pulses that do not cause irritation. In the figur the curve is called *excitation characteristic curve.* It is specific to differen muscles.

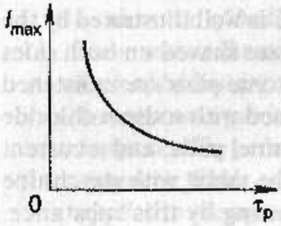


Fig. 19.2

Since a specific physiological action of an electric current depends on the pulse shape, in medicine, for stimulation of the central nervous system (electric sleep, electric anesthesia), neuromuscular system, cardiovascular syste (pacemakers, defibrillators), etc., currents with different time dependen are used.

Current with rectangular pulses with duration of pulses τ_p = 0.1–1 msec and the frequency range 5–150 Hz is used for electric sleep therap At τ_p = 0.8–3 ms the frequency range of 1–1.2 Hz, the current is used in implantable pacemakers. Currents with triangular pulses (fig. 19.3, a; τ_p = 1–1.5 ms, frequency 100 Hz), and currents with exponential shape, whos

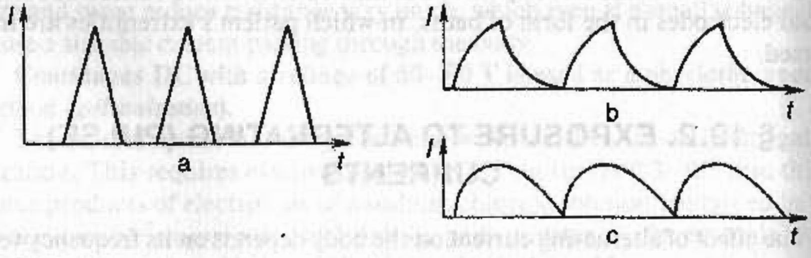


Fig. 19.3

pulses increase slowly and drop quite quickly (fig. 19.3, b; $\tau_p = 3-60$ ms, frequency 8–80 Hz) are used for muscle stimulation, particularly, in electro gymnastics. Diadynamic currents proposed by Bernard are used for different types of electrotherapy. Fig. 19.3, b, shows the shape of one of such pulse current types, the pulse recurrence rate is about 100 Hz.

At low, sonic, and ultrasonic frequencies, the effect of alternating (harmonic) current on the body is estimated by the following threshold values: the threshold of perceptible current and the threshold of freezing current.

The threshold of perceptible current is the lowest current whose irritating intensity can be felt by a human. This value depends on the place and area of contact of the body with the power supplied, current frequency, individual characteristics of the person (gender, age, specific features of his or her body). For homogeneous groups of subjects, the threshold of perceptible current meets the law of normal distribution with an average value of about 1 mA at a frequency of 50 Hz in men for the forearm-hand section. Fig. 19.4 (curve *1*) shows the dependence of average threshold of perceptible current for this group of subjects on the current frequency.

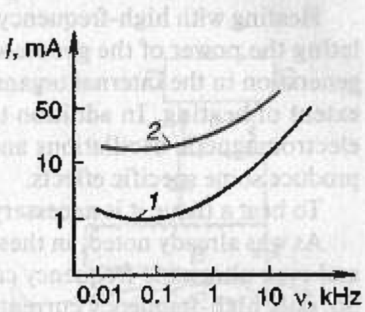


Fig. 19.4

If we increase the current intensity from the threshold of its perceptible value, it is possible to cause such a flexion of his or her joint, at which the person cannot unclench his or her hand, and free himself from the conductor, which is the source of power. The minimum intensity of such current is called the ***threshold of freezing current.*** Currents of lower intensity are ***releasing currents.*** The threshold of freezing current is a significant parameter: its excess can be fatal to humans. The values of the threshold of freezing current also meets the law of normal distribution. Fig. 19.4 (curve *2*) shows the dependence of the average value of freezing current threshold for the group of tested men on current frequency.

Acting on the heart, a current can cause ventricular fibrillation, which leads to death. The threshold of the current intensity causing fibrillation depends on the density of the current passing through the heart, the frequency and duration of the current action.

At frequencies more than approximately 500 kHz, displacement of ions becomes comparable with their displacement due to their molecular-thermal motion. That is why a current or electromagnetic wave will not cause an irritating effect. The main primary effect in this case is thermal impact. Therapeutic heating with high-frequency electromagnetic fields has a number of advantag over such a traditional and simple method as using a hot-water bottle.

Heating of internal organs with a hot-water bottle occurs due to the thermal conductivity of external tissues — the skin and subcutaneous fat. Hig frequency heating takes place due to heat generation in the internal pa of the body, i.e., it can be generated where it is necessary. The released he depends on dielectric permeability of tissues, their specific resistance, and the frequency of electromagnetic oscillations. Choosing the appropriate freque cy it is possible to implement a "thermoselective" action, i.e. preferential formation of heat in tissues and organs where required.

Heating with high-frequency oscillations is also convenient because, re lating the power of the generator, it is possible to control the power of h generation in the internal organs, and in certain treatments, one can dose the extent of heating. In addition to a thermal action, at their high frequenc electromagnetic oscillations and waves cause intramolecular processes that produce some specific effects.

To heat a tissue it is necessary to generate a high intensity current.

As was already noted, in these cases the direct current, as well a low, son and even ultrasonic frequency can provoke electrolysis and tissue destructi So, only high-frequency currents are used for heating (see § 18.9).

The current power consumed to heat the tissues is calculated by formu $P = I^2R$. Let us transform it assuming that the biological tissue is placed between two flat electrodes with area S located at distance l (as it is show in fig. 15.2).

Let the current density j be equal at all points of the tissue and coinci with the current density on the electrodes. Taking into account that $R = \rho l/S$ we obtain:

$$P = I^2R = j^2S^2\rho l/S = j\rho^2V, \qquad (19.1)$$

where $V = Sl$ is the tissue volume. Dividing (19.1) by this volume, we obtai the quantity of heat q released per 1 s in 1 m^3:

$$q = j\rho^2. \qquad (19.2)$$

As was to be expected, q depends on the current density and specific resistance of the tissue.

Transmission of a high frequency current through the tissue is used in physiotherapeutic treatments called diathermy and local darsonvalization.

Diathermy requires a current with frequency of about 1 MHz, weak damped oscillations and voltage 100–150 V; the current intensity is about several amperes. Since the skin, fat, bones, and muscles have the highest resistan they receive more heat. The least heating is in the organs rich in blood or lymph: the lungs, liver, lymph nodes. A disadvantage of diathermy is that a large number of heat released in the skin layer and subcutaneous tissue is wasted.

Recently, diathermy has been relinquished by clinical practice and replaced by other methods with a high-frequency effect. This is due to increased risk of diathermy: failure of the device, accidental arcing at the place of electrodes in direct two-pole contact of bimetallic object and sizeable current intensity can lead to tragic consequences.

For local darsonvalization, a current with a frequency of 100–400 kHz is used; the voltage is in tens of kilovolts, and the current intensity is small — 10–15 mA.

The patient P (fig. 19.5) receives the current supplied from a source of high-frequency oscillations S through vacuum or graphite-filled glass electrode E. There is no other electrode as the circuit is closed (the capacitor is shown with a dotted line) through the patient's body and the environment because there are displacement currents. The active factor is not only a high-frequency pulse current, but also an electrical discharge that occurs between the patient's skin and the electrode.

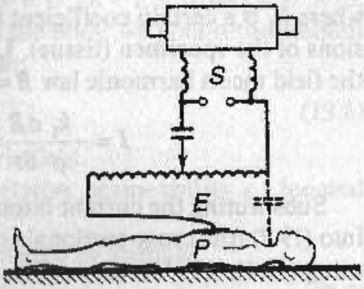


Fig. 19.5

High frequency currents are also used in surgery (*electrosurgery*). They make it possible to burn, fuse tissues (*diathermocoagulation*) or dissect them (*electrotomy*).

For diathermocoagulation, a current density of 6–10 mA/mm^2 is effective; as a result, the tissue temperature rises and the tissue coagulates. In electrotomy, the current density attains 40 mA/mm^2; hence, it is possible to cut a tissue using a sharp electrode (electric knife). Electrosurgery has certain advantages over conventional surgery.

§ 19.3. EXPOSURE TO ALTERNATING MAGNETIC FIELD

In § 17.4, it was shown that eddy currents occur in massive conducting bodies in an alternating magnetic field. These currents can be used to heat biological tissues and organs. This treatment method — ***inductothermy*** — has a number of advantages over the method described in § 19.2.

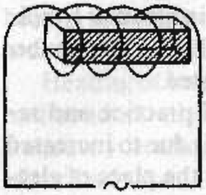


Fig. 19.6

Let us consider, on what factors the degree of heatin of tissues depends in inductothermy. A diagram of its effect is shown in fig. 19.6. From formulas (17.6) and (16.7), it is possible to write down the following approxima formulas for eddy currents:

$$I = \frac{S}{R}\frac{dB}{dt}.$$

Expressing R by the formula for resistance, we obtain:

$$I = -\frac{k_1}{\rho}\frac{dB}{dt}, \qquad (19.3)$$

where k_1 is a certain coefficient that takes into account the geometric dimensions of the specimen (tissue). Let us assume that the magnetic induction of the field meets harmonic law $B = B_{\max}\cos\omega t$, then

$$I = -\frac{k_1}{\rho}\frac{dB}{dt} = -B_{\max}\,\omega\sin\omega t. \qquad (19.4)$$

Substituting the current intensity instead of the current density from (19.3 into (19.2) (the cross-sectional area is taken into account in coefficient k_1) and using (19.4), we obtain:

$$q = \frac{k}{\rho^2}B_{\max}^2\omega^2\rho\sin^2\omega t = k\frac{\omega^2}{\rho}B_{\max}^2\omega^2\rho\sin^2\omega t, \qquad (19.5)$$

where k is a certain coefficient.

Thus, in inductothermy, the amount of heat released in the tissues, is proportional to the squares of the frequency and induction of the alternati magnetic field and inversely proportional to specific resistance. Therefore, tissues rich in blood vessels, such as muscles, will be heated more than tissues like fat. Usually, a local impact of an alternating magnetic field is used in inductothermy, using spirals or flat rolled cables.

Eddy current therapy is also possible in general darsonvalization. In this case, the patient is placed in a solenoid cage, the turns of which are fed with high-frequency pulse current.

§ 19.4. EXPOSURE TO ALTERNATING ELECTRIC FIELD

In tissues located in an alternating electric field (see diagram in fig. 19.7) displacement and conduction currents are generated. Usually, electric fields

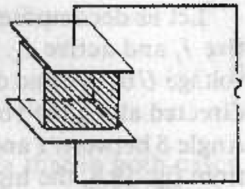


Fig. 19.7

of ultrahigh frequency are used for this purpose, so the corresponding physiotherapeutic method is called UHF *therapy*. To estimate the efficiency of the UHF field effect, it is necessary to calculate the number of heat released in conductors and dielectrics.

Let a body conducting an electric current be placed into an alternating electric field. In this case, the electrodes do not touch the body. Therefore, it is expedient to express the amount of heat released in the body not by the current density on the electrodes [see (19.2)], but by intensity E of the electric field in the conducting body.

Let us do quite simple transformations: $P = U^2/R = E^2 l^2 S/(\rho l) = E^2 Sl/\rho$. Having divided this equation by volume Sl of the body, we obtain the amount of heat released during 1 s in 1 m^3 of the tissue:

$$q = P/(Sl) = E^2/\rho, \qquad (19.6)$$

where E is the effective intensity[1] of the electric field.

Now, let us consider a dielectric with relative permeability ε_r, located in an alternating electric field.

In an AC circuit, the average power is expressed by formula:

$$P = (U_{\max} I_{\max}/2)\cos\varphi = U_{ef} I_{ef} \cos\varphi, \qquad (19.7)$$

where φ is the phase difference between the current intensity and voltage. If we apply formula (19.7) to a capacitor (see fig. 18.6, a), then taking into account that $\varphi = \pi/2$, we obtain:

$$P = 0.$$

In this example (see the vector diagram in fig. 18.6, b), the power is not absorbed in the capacitor, and the displacement current is fully reactive. In a real dielectric, a low conduction current and periodic change in polarization cause absorption of the supplied electric power, the dielectric heats up, which partially consumes the alternating electric field energy consumed, i.e. there are dielectric losses.

As we can see from formula (19.7), losses in the dielectric mean that there will be a phase shift between the current intensity and voltage; the shift is equal to $\varphi \neq \pi/2$ (fig. 19.8).

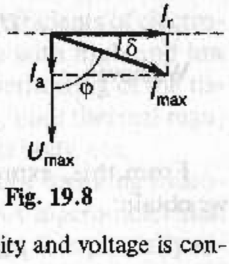


Fig. 19.8

---

[1] An effective electric field intensity similar to current intensity and voltage is connected with the maximum value by ratio $E = E_{\max}/\sqrt{2}$.

Let us decompose current vector I into a sum of two components: reactive I_r and active I_a. The reactive component is phase shifted relative to voltage U by $\pi/2$ and does not cause dielectric losses; the active component is directed along the voltage vector, and this is what causes dielectric losses. Angle δ between I and I_r is called an ***angle of dielectric losses***. As we can see from fig. 19.8, the higher the angle, the higher the active component of the current intensity is. In practice, both reactive and active components of current intensity are connected through the tangent of the dielectric losses angle:

$$\frac{I_a}{I_r} = \tan\delta,\ \ I_a = I_r \tan\delta. \tag{19.8}$$

From fig. 19.8, we can see that $I_a = I_{max}\cos\varphi$; comparing it with (19.8) we obtain:

$$I_{max}\cos\varphi = I_r \tan\delta. \tag{19.9}$$

Taking into account (19.9), we transform the formula for the power (19.7):

$$P = \frac{U_{max}}{2} I_r \tan\delta. \tag{19.10}$$

The amplitude of the reactive component of current intensity I_r is the actual amplitude of current intensity which is an amplitude amplitude corresponding to an ideal capacitor [see (18.28)]. Therefore,

$$I_r = U_{max} C\omega. \tag{19.11}$$

Substituting (19.11) into (19.10) and expanding the expression for the capacity of a flat capacitor, we obtain average power:

$$P = \frac{U_{max}^2}{2}\frac{\varepsilon_r\varepsilon_0 S}{l}\omega\tan\delta. \tag{19.12}$$

Instead of voltage amplitude U_{max}, we use its effective value:

$$U_{ef} : U_{ef} = U_{max}/\sqrt{2} \quad \text{or} \quad U_{ef}^2 = U_{max}^2/2.$$

We have:

$$P = U_{ef}^2\omega(\varepsilon_r\varepsilon_0 S/l)\tan\delta.$$

From this, expressing the voltage through the electric field intensity we obtain:

$$P = E^2 l^2 \omega(\varepsilon_r\varepsilon_0 S/l)\tan\delta = \omega E^2 \varepsilon_r \varepsilon_0 \tan\delta S l.$$

Having divided this equality by the dielectric volume Sl, we obtain:

$$q = \frac{P}{Sl} = \omega E^2 \varepsilon_r \varepsilon_0 \tan\delta. \qquad (19.13)$$

(by E, we mean the effective electric field intensity).

Comparing formulas (19.6) and (19.13) we can see that in both cases, the amount of released of heat is proportional to the square of the effective electric field intensity. It also depends on properties of the medium, and for the dielectric, on the actual frequency of the field.

In Russia, UHF devices work at a frequency of 40.58 MHz. If the currents have such a frequency, the dielectric tissues of the body are heated more intensively than conducting tissues.

§ 19.5. EXPOSURE TO ELECTROMAGNETIC WAVES

Physiotherapeutic methods based on electromagnetic waves of UHF range, have two names depending on the wavelength: *microwave therapy* (frequency of 2.375 MHz, wavelength of 12.6 cm) and *UHF wave therapy*, i.e., therapy of ultra high frequency waves (frequency of 460 MHz, wave length of 65.2 cm).

The theory of thermal effect of UHF fields on biological objects is the most developed at present. The electromagnetic wave polarizes the molecules of the material and reorients them as electric dipoles periodically. In addition, the electromagnetic wave affects ions of biological systems and causes an alternating current of conductivity. Thus, in a material placed in an electromagnet field, there are both displacement and conduction currents. All these factors lead to heating of the material. Currents of displacement caused by reorientation of water molecules are of great importance. In this connection, the maximum absorption of UHF waves energy takes place in such tissues as muscles and blood, while osseous and adipose tissues have less water, so they are heated less.

At the interfaces of media with different absorption coefficients of electromagnetic waves, for example, at the interface of tissues with high and low water content, standing waves can occur causing local overheating of the tissues. Tissues with insufficient blood supply and, therefore, poor thermal regulation, are the most prone to overheating: the lens, vitreous body, etc.

Electromagnetic waves can affect biological processes by breaking hydrogen bonds and changing the orientation of DNA and RNA macromolecules.

When an electromagnetic wave hits a part of the body, it is partially reflected from the surface of the skin. The degree of reflection depends on the difference between dielectric permeability of the air and biological tissues. If irradiation with electromagnetic waves is carried out remotely (at a distanc then up to 75% of the energy of electromagnetic waves can be reflected. In this case, it is impossible to estimate the energy absorbed by the patient per a time unit, by the power generated by the emitter. In case of contact irradiation with electromagnetic waves (the emitter is in contact with the surface irradiate the power generated corresponds to the power received by body tissues.

The depth of electromagnetic wave penetration into biological tissues depends on the ability of these tissues to absorb wave energy, which, in turn, is determined by both the structure of tissues (mainly water content) and the frequency of electromagnetic waves. Thus, centimeter wavelength electromagnetic waves used in physiotherapy pass into muscles, skin, biological fluids to a depth of about 2 cm, and into fat, bones, the waves pass to a depth of about 10 cm. For UHF waves, these values are approximately 2 times higher.

Taking into account the complex composition of tissues it is conventiona considered that in case of UHF wave therapy, the depth of penetration of electromagnetic waves is equal to 3–5 cm from the surface of the body, and for UHF therapy, it is up to 9 cm.

Section 5

GENERAL AND MEDICAL ELECTRONICS

The concept of electronics is widespread at present. As a technical science, electronics is primarily based on the achievements of physics. We can definitely say that without electronic equipment, it is nowadays impossible to diagnose diseases nor to administer their effective treatment. In this section, we deal only with some other important aspects of general and medical electronics as well as the most essential medical electronic equipment. Some instruments and devices of medical electronics are considered in other sections.

Chapter 20

Content of general and medical electronics

In this chapter, along with presentation of the content of electronics, we consider problems of electrical safety and reliability of medical electronic equipment.

§ 20.1. ELECTRONICS AND SOME LINES OF ITS DEVELOPMENT

Physics, like any other science, was developing and is developing in connection with the needs of society; its progress is motivated by practical tasks. In turn, development of physics contributes to solution of practical tasks including technical problems. For example, due to achievements in the field of electromagnetic phenomena research, corresponding branches of technology have developed rapidly: electrical and radio engineering. Gradually, many sections of radio-engineering start to be called ***radio electronics*** **or simply** ***electronics.***

The term electronics is conventional to a great extent, it is difficult to give it a clear definition. Maybe, the most correct understanding of electronics is as a field of science and technology, which deals with operation and application of vacuum, ans solid-state ionic devices.

In the broad sense of the word, electronics (general electronics) can be divided either into groups according to its application, according to the class of devices used, or the category of theoretical questions. Therefore, there are physical electronics, referring to the section of physics dealing with electrical conductance of bodies, contact, thermoelectronic phenomena; technical electronics describes the structure of devices and instruments and their connection diagrams; semiconductor electronics deals with application of semiconductor devices, etc.

Sometimes, all electronics is divided into three major areas: vacuum electronics covering the design and application of vacuum devices (electronic tubes, photoelectric devices, X-ray tubes, gas-discharge devices); semiconductor electronics, which covers the design and application of semiconductor devices, including integrated circuits; quantum electronics is a specific section of electronics related to lasers and masers.

All these examples, on the one hand, give a general idea of the content of electronics; on the other hand, they emphasize indefiniteness of the limits of electronics once again.

Electronics is a very dynamic section of science and technology. Based on new effects (phenomena), electronic devices are developed, among them, those used in biology and medicine. As an illustration, let us consider the ***Josephson effect.***

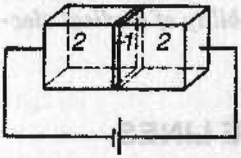


Fig. 20.1

In 1962, the English physicist B. Josephson predicted that through a very thin layer of dielectric *1* (about 1 nm thick) located between two superconductors *2*, an electric current could flow (fig. 20.1). Shortly after, such a phenomenon was indeed discovered. When the current intensity was less than a certain critical value, the dielectric did not generate a voltage drop; but if the current intensity exceeded the critical value, the dielectric generated a voltage drop, and the contact superconductor — dielectric — superconductor emitted electromagnetic waves.

The critical value of current intensity is sensitive to external magnetic fields. So a device like the one shown in fig. 20.1, can be used for accurate measurement of a weak magnetic field intensity, whose values can be about 10^{-7} A/m. The Josephson effect recently began to be used for measurement of the magnetic field induction of cardiac biocurrents (see § 16.8).

Any technical device, including radio or electronic one tend to be modernized and made more reliable, consuming less energy, compact, etc. Nevertheless, there are difficulties: for example, compactness of a device can reduce its reliability, etc.

Let us concentrate on one side of the problem: the trend towards reducing the dimensions, or miniaturization, of electronic devices and circuit elements. Even at the time when electronics used only electronic tubes there was an effort to make these tubes, as well as passive circuit elements (resistors, inductors, capacitors) as small as possible. Later, printed circuits started to be introduced; they had an advantage over usual circuits in size and, in addition, made it possible to mechanize the process of mounting circuits. These trends led to the fact that in the early 1950s, it was possible to design electronic devices where an area of 1 cm^3 of which accommodated 0.5 elements.

In miniaturization of electronic devices, a significant improvement was introduction of semiconductor diodes and triodes, which permitted attaining the density of electronic devices to 2–3 elements in 1 cm^3.

The next stage of miniaturization of electronics, which is developing nowadays, is developing ***integrated circuits***. These are micro-miniature electron devices, in which all components (or their parts) are structurally and integrally related and connected electrically.

There are two main types of integrated circuits: semiconductors and film integrated circuits.

Semiconductor integrated circuits are made from high-purity semiconductors. Thermal treatment, diffusion and other procedures change the crystal lattice of a semiconductor so that its separate regions become different elements of the circuit. This makes it possible to design a circuit equivalent for a radio-engineering unit consisting of 100 or more components from a plate of about 1 mm^2. In integrated circuits, p–n-junctions are usually used as resistors and capacitors.

Fig. 20.2 shows a semiconductor integrated circuit: a is its cross-section; b is its wiring diagram. Different shading depicts various semiconductor or conductive materials or materials with different additives (A is the capacitor, B is the triode, C is the resistor, 1–5 are the corresponding points of the diagram).

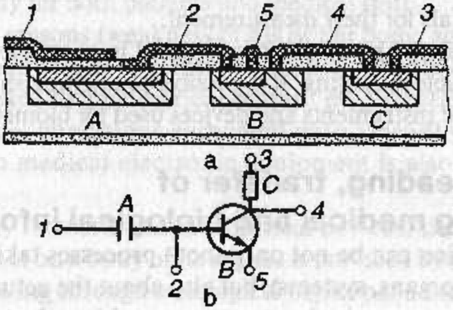


Fig. 20.2

Film integrated circuits are manufactured by deposition of various materials on corresponding substrates in vacuum.

There are also ***hybrid integrated circuits***, which are a combination of semiconductor and film integrated circuits.

The dimensions of individual elements of integrated circuits are about 0.5–10 μm, so the smallest specks of dust, motes, etc. can affect their work. Therefore, manufacturing of integrated circuits should be done in conditions of heightened purity of the environment.

Integrated circuits containing of more than 100 elements are called ***large-scale integrated circuits*** (LSIC).

Developing integrated circuits, miniaturization of electronic devices is one of the main directions of modern electronics development.

§ 20.2. MEDICAL ELECTRONICS. MAIN GROUPS OF MEDICAL ELECTRONIC INSTRUMENTS AND DEVICES

Electronics is an applied branch of knowledge. A common application of electronic devices is connected with diagnostics and treatment of diseases. Sections of electronics, dealing with specialties of electronic systems used [illegible] solve biomedical problems, as well as the structure of appropriate equipment are called ***medical electronics***.

Medical electronics is based on data from physics, mathematics, engineering, medicine, biology, physiology, and other sciences; it includes biological and physiological electronics.

Applications of electronics in medicine are diverse because this is a constantly expanding area.

Nowadays, many traditionally non-electrical characteristics like temperature, body displacement, biochemical indicators, etc. tend to be converted into electrical signals for their measurement.

Information carried by an electrical signal is convenient for transfer over a distance and reliable recording. It is possible to distinguish the following main groups of electronic instruments and devices used for biomedical purposes:

Devices for reading, transfer of and recording medical and biological information

Such information can be not only about processes taking place in a body (biological tissue, organs, systems), but also about the actual state of the environment (sanitary purpose), about processes taking place in prostheses, etc. This includes most diagnostic equipment: ballistocardiography, phonocardiographs, rheography, etc. With respect to radio-engineering, most such devices are characterized by the presence of amplifiers of electrical signals.

This group can include electro-medical equipment for laboratory studies such as pH-meter.

Electronic devices providing exposure to various physical factors

Electronic devices that provide exposure to various physical factors (ultrasound, electric current, electromagnetic fields, etc.) for therapeutic purposes: devices of microwave therapy, devices for electro-surgery, cardiac pacemakers, etc.

From the physical viewpoint, these devices are generators of various electrical signals.

Cybernetic electronic devices

Cybernetic electronic devices:

a) computers for processing, storage, and automatic analysis of medical and biological information;

b) devices for controlling vital functions and automatic regulation of environmental conditions for a human;

c) electronic modeling of biological processes, etc.

Application of electronic medical instruments and devices increases the efficiency of diagnostics and treatment as well as the work productivity of medical staff.

§ 20.3. ELECTRICAL SAFETY OF MEDICAL EQUIPMENT

An important problem related to the use of electronic medical equipment is its electrical safety for both patients and medical staff.

Due to various reasons (weakness of his or her body, action of anesthesia, unconscious state, presence of electrodes on his or her body, i.e, direct connection of the patient to an electrical circuit, etc.) a patient is in especially dangerous electrical conditions compared with a healthy person. Medical staff working with medical electronic equipment is also at risk of electric shock.

In an electricity network and in technical devices, the voltage is usually pre-set, but the effect on a body or its organs is provided by an electric current, that is a charge flowing through a biological object per a time unit.

Between two terminals (electrodes), the resistance of a human body consists of the resistance of internal tissues, viscera, and the resistance of the skin (fig. 20.3). Resistance R_{int} of internal parts of the body does not depend much on the general state of the person; in calculations, it is assumed that $R_{\text{int}} = 1$ kOhm in the palm–foot path. The resistance of skin R_s exceeds the resistance of viscera greatly and depends essentially on internal and external causes (sweating, humidity). In addition, the skin has different thickness on different parts of the body and, therefore, different resistance. So (taking into account the variability of human skin resistance) it is generally not taken into account at all; it is considered that $I = U/R_{\text{int}} = U/1000$ Ohm. For example, $I = 220/1000$ A = 220 mA at $U = 220$ V. In fact, the skin has resistance, hence, in a real situation at voltage

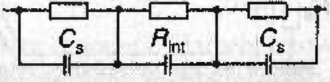


Fig. 20.3

of 220 V, the current is less than 220 mA. It is clear that when using electron medical equipment all possible safety measures should be provided.

The main and most important prerequisite is the following: to rule out a contact with components of equipment, where there is voltage.

For this purpose, parts of instruments and devices, where there is voltag are first insulated one from another and from the frame of the equipmen Insulation, which is required for such a purpose, is called operational insulation. Holes in the frame should rule out accidental penetration inside and contact with the inner parts of the equipment with fingers, jewelry chains, etc. Nevertheless, even if components of the equipment where there is voltage are protected from a contact, it does not yet ensure full safety for at least two reasons.

First, between the internal parts of the equipment and its frame, whatev insulation is, the resistance of devices and instruments against an alternatin current is not boundless. The resistance between the electrical wiring and the earth is not boundless either. Therefore, when a person touches the frame of equipment, a certain current will pass through the human body; this is called a *leakage current*.

Second, it is possible that due to deterioration of the working insulatio (growing old, humidity of the ambient air) the electric circuit closes on the internal part of the equipment and the frame referred to as as fault to fram and the external, touchable part of the equipment frame will carry voltage.

In both cases, one should take measures to prevent electric shock in peopl touching the frame of the instrument or device. Let us consider these problems in more detail.

The intensity of current leakage onto the frame depends like any conduction current on the voltage and resistance of the circuit. A circuit with leakag current is represented schematically in fig. 20.4. In this, *1* is the frame of device, with a transformer inside; *2* is primary winding of the transformer whic is connected to the power point of circuit *3*. Secondary winding *4* of the transformer is connected to the working section of the equipment (not show in the figure). The circuit, irrespectiv of the presence or absence of earth, has always certain conductivity relative to the earth. Active (ohmic) resistance R_e of insulation and earth, as well as capacity C_e of circuit conductors and the

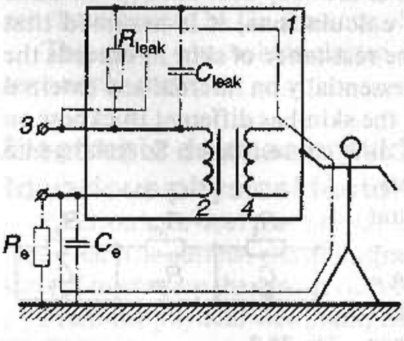


Fig. 20.4

earth determines the conductivity. The electrical conductivity between the circuit and the frame depends, respectively, on the ohmic resistance of the working insulation and capacitance between the internal parts of the device carrying voltage and the frame, that is on R_{leak} and C_{leak}. All these elements are shown in fig. 20.4 bydotted lines because they are distributed parameters and do not present resistors or capacitors. The dash-dotted line in the figure shows the path of the leakage current passing through the human body touching the frame of the instrument or device.

If the patient's circuit is insulated from the frame then leakage current onto the patient is distinguished separately.

Since the leakage current affects the safety of operation of medical equipment very much, for designing and manufacturing of these articles, they take into account the permissible intensity of this current both for normal operation of instruments and devices and in case of a single failure. A ***single failure*** means a failure of one of the means of protection against electric shock. According to electrical safety regulations, a single failure should not create a direct danger to a person. Permissible intensities of leakage currents are distinguished by types of electrical medical articles depending on their degree of protection against electric shock. There are the following four types:

1. H means articles with a normal degree of protection; this protection is equivalent to the protection of household appliances.
2. B means articles with a heightened degree of protection.
3. BF means articles with a higher degree of protection and insulated working part;
4. CF means articles with the highest degree of protection and insulated working part. This type includes compulsorily, in particular, articles with the working part having electrical contact with a heart. CF type articles should have special marking (fig. 20.5, a).

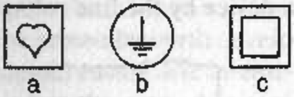


Fig. 20.5

Table 20.1 shows permissible leakage current intensity (mA) for articles type N, H, BF, and CF.

In case of fault to frame, touchable (external) parts of a device carry voltage. In this case, at abnormal working conditions of articles, it is necessary to foresee all possible protection against electric shock. Such basic protective measures include earthing and zeroing. To understand the physical side of these measures, it is necessary to know how electric medical equipment is connected to a three-phase system.

For technical solution of the problem of the most economic filament transfer of AC, at the end of the 19th century, the Russian engineer M.O. Dolivo-Dobrovolsky proposed three-phase electric power system (three-phase current).

Table 20.1

	Article type					
	H	C	B, BF		CF	
Leakage current type	Normal state	Single failure	Normal state	Single failure	Normal state	Single failure
Fault to frame	0.25	0.5	0.1	0.5	0.01	0.5
Onto the patient	—	–	0.1	0.5	0.01	0.05

A variant of this system is shown in fig. 20.6: *1* means phase windings of one generator; in them, alternating voltage is induced; *2* means loads (consumers); *3* means line conductors (they connect the generator with consumers). To ensure that the consumer in one circuit of a three-phase circuit does not affect the mode of operation of another circuit, it is expedient to connect a neutral (zero) filament *4*. Voltages U_l between the line conductors are called *linear*; voltages between linear and neutral conductors are referred to as *phase* voltages (U_{ph}). The relationship between phase and line voltages is the following:

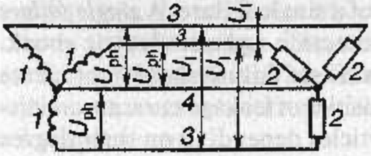


Fig. 20.6

$$U_l = \sqrt{3}U_{ph} \approx 1{,}73U_{ph}.$$

Electro-medical equipment is usually connected as a single-phase load to line or phase voltage. Fig. 20.7 shows the power supply of an instrument or a device by the line voltage; the neutral wire is insulated. To simplify, we assume that the line conductors have perfect insulation and the neutral wire has resistance R_{ins} relative to the ground (shown with the dotted line). If it was not protective earth R_e, in case of fault to frame and touching the frame by a person, he or she would carry voltage. The dot-dash line shows the circuit to which the person would be connected. It is clear from the figure that voltage U_{ph} is redistributed between resistance R_p of the person's body, including his or her earthing, and resistance R_{ins}. If, for instance, $R_p = 0.5R_{ins}$ and $U_{ph} = 220$ V, then the person can have

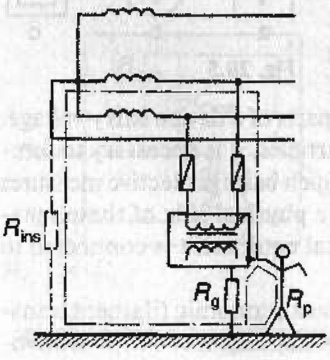


Fig. 20.7

20/3 V ≈ 75 V. For protection of the person, in this case, it is necessary to earth the frame. Resistance R_g of the earth is connected in parallel to R_p. Since R_e is low (should not exceed 4 Ohm), then $R_{ins} \gg R_e$; actually on this resistor, and, therefore, on the person, the voltage would be quite negligible.

It is important to note that due to sizeable resistance R_{ins}, a fault to frame will not cause an emergency current, which is sufficient to trigger the safety device, so this abnormality can remain unnoticed by the staff. Nevertheless, if nearby there is a device (instrument) with a fault to frame from another line conductor (another phase), then line voltage will occur between the frames of two devices. Simultaneous touch to such frames is very dangerous.

Nowadays, three-phase circuits with an earthed neutral are mostly common. In this case, protective earth is ineffective. In fact, with good earthing (fig. 20.7), R_{ins} is low, for instance, $R_{ins} \approx R_g$. Voltage U_{ph} will be halved between these resistances, and there will be voltage between the frame and earth, which is equal to $0.5U_{ph}$. It is dangerous for a human. Most likely, the protective device will work during the spark-over, but this can happen not immediately or even never at all, if the emergency current intensity is insufficient. In order for the protective device to work, another type of protection is used — protective zeroing, at which the frame of equipment is connected by conductors to the neutral conductor of the power supply (fig. 20.8). In case of fault to frame, short circuit occurs shown (by the dash-dot line), the protective device works and equipment is cut off from the voltage source. Since there is always a probability of neutral conductor rupture, the neutral conductor is earthed in several places.

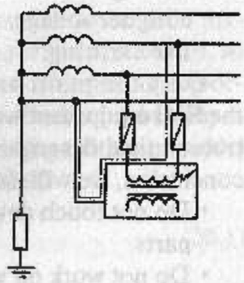


Fig. 20.8

Summing up the above we note once again that the protective earth or zero in installations with an isolated neutral conductor, should ensure safe intensity value of current passing through the human body when the network is connected with the earthed parts of the equipment. In addition, in installations with an earthed neutral conductor, grounds or zeros should cause automatic disconnection of the equipment from the electric power supply.

However, not all electro-medical equipment is reliably protected by earthing or neutralizing. Depending on the means of additional protection against electric shock from a power supply, equipment is divided into the following four classes:

- I is articles, which in addition to the main insulation require joint connection of supply voltage and earthing (zeroing) of touchable metal parts. For example, this can be done with three conductor power supply cable and three contact plug. Two conductors of the cable are used to pow supply and the third is an earth. When you insert the plug into for sock the earth is connected before the voltage.

- I is articles, which differ from articles of class I only by having a separ clamp (terminal) on their touchable metal parts in order to connect the to an external earthing (zeroing) device. Fig. 20.5, b, shows the pla of protective earth (zero). Use of articles of class 0I is temporary; the articles should be later replaced by the corresponding class I products
- II is articles, which in addition to operational insulation, have suppl mentary insulation. It is possible to have reinforced insulation inste of operational or supplementary. Equipment of this class does not hav protective earth facilities. Fig. 20.5, a, shows the lead-in of a power cord (or cable) for articles of this class.
- III is for articles that are designed to be powered from an isolated pow supply with an alternating voltage not higher than 24 V or a DC volta not higher than 50 V. They do not have external or internal circuits with a higher voltage. Articles of this class also do not have devices for prote-tive earthing.

Only the main problems of electrical safety for working with electr-medical equipment were considered above. Since it is difficult to give an electrotechnical description of various situations that can lead to an acciden in conclusion, we will limit ourselves to certain general instructions.

- Do not touch devices with two bare hands simultaneously and your bod parts.
- Do not work on wet, damp floor, on the ground.
- Do not touch pipes (gas, water, heating), metal constructions whe working with electrical equipment.
- Do not touch metal parts of two devices at the same time.

When administering procedures using electrodes applied on a patient, it i difficult to foresee many variants of generating electrically dangerous situations (patient touching a heating battery, gas and water pipes and valves, shor circuit through the frame of adjacent equipment, etc.). Therefore, you should strictly follow the instructions on carrying out this procedure without any deviations from the rules,

§ 20.4. RELIABILITY OF MEDICAL EQUIPMENT

Medical equipment should function normally. This requirement, howeve is not always met; more exactly, such a requirement cannot be fulfilled for any given duration unless special measures are taken.

A physician using medical equipment should have an idea of the probability of failure occurrence, that is, the probability of wear and tear of the device (apparatus) as a whole or its components, of an increase or decrease in permissible parameters. A device that does not meet the specifications becomes inoperative. Having been repaired the device can be operative again. In many cases, it is sufficient to replace a tube or resistor to make the article function normally again, but it can happen that the equipment is so out of date and used that it is not economically expedient to repair (restore) it. In this connection, the medical staff should have an idea of the maintainability of the equipment and the durability of its parts.

The ability of an article to work under specified operating conditions well and to keep its performance for a given interval of time is characterized by the general term *reliability*.

For medical equipment, the problem of reliability is of special importance, since unserviceable instruments and devices can bring about not only economic losses, but also patients' death.

The ability of equipment to provide trouble-free operation depends on many factors; it is almost impossible to take into account the effect of them all, so the quantitative estimate of reliability is probabilistic. For example, an important parameter is the *probability of trouble-free operation*. **Such a probability is estimated experimentally by the ratio of number N of operating (without trouble) articles during time t to the total number N_0 of articles tested:**

$$P(t) = \frac{N(t)}{N_0}. \tag{20.1}$$

This parameter estimates the ability of an article to remain operable for a given time interval. Another qualitative indicator of reliability is *breakdown rate* $\lambda(t)$. This indicator is equal to the ratio of the number of failures dN to the product of time dt by total number N of operating elements:

$$\lambda = -\frac{dN}{N dt}. \tag{20.2}$$

The negative sign is here because $dN < 0$ as the number of operating articles decreases over time.

Function $\lambda(t)$ may vary. The most characteristic shape of its curve is shown in fig. 20.9. Three domains are noticeable in it: I is adjusting time interval, when defective elements of an article become disabled; latent troubles that that appeared in the process of manufacturing manifest themselves. The breakdown rate can be quite high. II is the time interval of normal operation, the breakdown rate can be constant for a long time. Normal operation

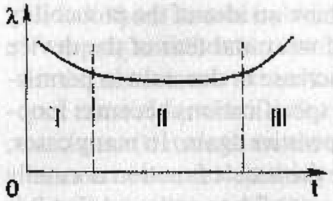


Fig. 20.9

of the equipment should be planned for this time period. III is the aging time interval: the breakdown rate increases over time due to wear of materials and components.

Concerning medical specialists, it would be interesting to note that the time dependence curve of the parameter characterizing human mortality rate is of approximately the same shape. An intensive rate is more typical of infants (time interval I) and old people (time interval III).

There is a certain relation between trouble-free operation probability P and breakdown rate λ. Let us determine the relation for cases if $\lambda = \text{const}$ (time interval II). We will write a differential equation (20.2) separating the variables on opposite sides of the equation:

$$\frac{\mathrm{d}N}{N} = -\lambda\,\mathrm{d}t. \tag{20.3}$$

Integrating and substituting lower limits (initial number N_0 of tested articles and time $t = 0$) and upper limits (number N of trouble-free operating articles at the time point t), we obtain:

$$\int_{N_0}^{N}\frac{\mathrm{d}N}{N} = -\lambda\int_{0}^{t}\mathrm{d}t,\quad \ln\frac{N}{N_0} = -\lambda t,\quad \frac{N}{N_0} = e^{-\lambda t}. \tag{20.4}$$

Comparing (20.4) with (20.1), we have:

$$P(t) = e^{-\lambda t}. \tag{20.5}$$

Thus, at a constant breakdown rate, we end up with an exponential law of change over time of the trouble-free operation probability[1]. This law can be used to assess the reliability of equipment.

Depending on possible aftermath of failure during operation, medical devices are subdivided into the following four classes:

- A is articles, failure of which implies a direct danger to the life of the patients or staff. The probability of trouble-free operation of articles of this class should be not less than 0.99 during the operating time between planned preventive maintenances, and for articles, which are not subject to maintenance (repairing, checking) — during the service life established for them. Articles of this class include instruments for monitoring vital functions of a patient, artificial respiration and blood circulation devices, etc.

---

[1] It would be interesting and useful for the reader to compare equation (20.4) with another statistical equation, i.e. the basic law of radioactive decay (32.9).

- B is articles whose failure causes distortion of information about the actual state of the body or the environment, which does not lead to a direct threat to the life of a patient or staff or necessitates immediate use of a functionally analogous article that is in the standby mode. The probability of trouble-free operation of this class articles should be at least 0.8. Such articles include systems that monitor patients, devices for stimulation of cardiac activity, etc.
- C are articles whose failure reduces the efficiency of diagnosis and treatment or delays it in non-critical situations, increases the load on the medical or service staff or generates merely material damage. The mean time between failures (MTBF) of repaired articles of this class and the failure life of irreparable articles should not be less than twice the mean time between failures or the calendar time between scheduled preventive maintenance works. For articles not subject to maintenance, the MTBF should not be less than the warranty time between failures or the warranty period of operation at the average intensity of use of the articles. This class includes most diagnostic and physiotherapy equipment, tools, etc.
- D is articles that do not contain potential failure parts. Medical electrical equipment does not belong to this class.

Medical specialists should be know that the concept of reliability can be quite conventionally applied to the human body if we consider a disease as a loss of operability, treatment as repairing, and prevention as measures that contribute to enhancing reliability (maintenance). However, the body is a complex system and such a technical approach can be only partial taking into account feedbacks and regulation processes. This is the cybernetic approach to be exact (see Section 1).

Chapter 21

System of obtaining biomedical information

Any biomedical research involves obtaining and recording relevant information. Despite the variety of devices and methods used for this purpose, it is possible to specify their general schemes and principles of operation. The problems discussed in this chapter are partially related to cybernetics.

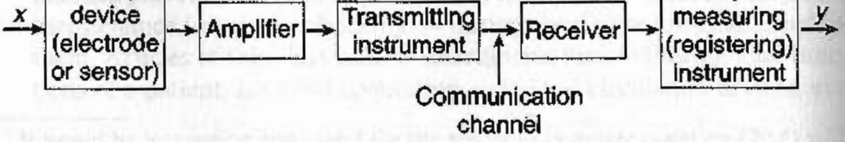
§ 21.1. STRUCTURAL DIAGRAM OF READING, TRANSMITTING AND RECORDING BIOMEDICAL INFORMATION

In order to obtain and record information about the state and paramete of a biomedical system, it is necessary to have a set of devices.

The primary element of this system is a sensor of a measuring instrumen called *reader*. It definitely contacts or interacts with the system itself, the othe elements are usually separated from the biomedical system, parts of the measurement system can sometimes be even remote at sizeable distances from the object of measurement.

A block diagram of the measuring chain is shown in fig. 21.1. This diagra is universal and reflects all kinds of real systems used in medicine for diagnosis and research. In medical electronic devices, the sensor either emits an electrical signal directly or modifies it under influence of a biological syste Thus, a reader converts information of biomedical and physiological conte into a signal of an electronic device. There are two types of readers in medica electric equipment: electrodes and sensors.

The final element of a measurement chain is a measuring instrument, which displays or registers information about a biological system in a form accessible for direct perception by an observer.



x
Pickup device (electrode or sensor)
Amplifier
Transmitting instrument
Receiver
Output measuring (registering) instrument
y
Communication channel

Fig. 21.1

In many cases, between the reader and measuring instrument, there are elements that amplify the initial signal (see Chapter 22) and transmit it over a distance.

In the block diagram, x means a certain measurable parameter of a biological system, such as the blood pressure. Letter y denotes the output value, for example, current intensity (mA) on the measuring instrument or an offset of the recorder (mm) on the paper of the recording device. For quantification we should know the dependence $y = f(x)$.

§ 21.2. ELECTRODES READING BIOELECTRICAL SIGNAL

Electrodes are special shape conductors connecting a measurement chain with a biological system.

In diagnostics, electrodes are used not only to read an electrical signal, but also to generate an external electromagnetic field, for example, in rheography. In medicine, electrodes are also used to provide exposure to electromagnetic field for therapeutic purposes and electrical stimulation.

There are certain prerequisites for electrodes: they should be easy to apply and remove, have high stability of their electric parameters, be service-strong, generate no interference, not irritate a biological tissue, etc.

An important physical problem relating to electrodes for reading bioelectrical signals is minimization of loss of useful information, especially on the electrode–skin transition resistance. An equivalent electric circuit diagram including a biological system and electrodes is shown in fig. 21.2 ($\mathcal{E}_{bp}$ is the emf of biopotential source; r is the resistance of internal tissues of biological system; R is the resistance of the skin and electrodes contacting with it; R_{inp} is input resistance of the amplifier of biopotentials). Based on Ohm's law, assuming that the current intensity on all segments of the circuit is equal, we obtain

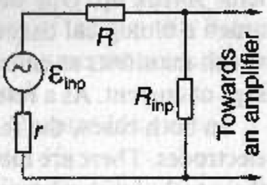


Fig. 21.2

$$\mathcal{E}_{bp} = Ir + IR + IR_{inp} = IR_i + IR_{inp} \quad (R_i = r + R). \tag{21.1}$$

At the amplifier input, we can conventionally designate the voltage drop as useful since the amplifier increases exactly this part of the emf of the source. Inside the biological system and on the electrod–skin system, from this viewpoint, voltage drop Ir and IR is useless. Since $\mathcal{E}_{bp}$ is preset and it is impossible to influence a decrease in Ir, then it is possible to increase IR_{inp} only decreasing R and first decreasing the resistance of the skin–electrode contact.

To reduce the intermediate resistance electrode–skin, they try to increa the conductivity of the medium between the electrode and the skin usin gauze bandages moistened with physiological solution or electrically conductive paste. This resistance can be reduced increasing the electrode–sk contact area, i.e. increasing the electrode dimension; nevertheless, the electrode will cover several equipotential surfaces (see, for example, fig. 14.1[illegible]), hence, the true picture of the electric field will be distorted.

By their application, electrodes for reading of a bioelectrical signal belon to the following groups:

1) for short-term use in surgery of functional diagnostics, for exampl for single reading electrocardiogram;
2) for long-term use, such as continuous monitoring of critical patients in intensive care units;
3) for use on mobile people being monitored, e.g., in sport or space medicine;
4) for emergency situations, such as in an ambulance conditions.

It is clear that in all cases electrode application has its specifics physiolog cal solution can dry up and the resistance will change if the observation of bioelectrical signals is long-term; with a patient in unconscious state it is more reliable to use needle electrodes, and the like.

When using electrodes in electro-physiological studies, two specific problems spring up. One of them is generation of galvanic emf when electrod touch a biological tissue. The other is electrolytic polarization of electrod which manifests as emission of reaction products on the electrodes during passage of current. As a result, there is an emf that is opposite to the main em

In both cases, the resulting emf distorts useful bioelectrical signal read by electrodes. There are methods to reduce or eliminate such unwanted effects, but these techniques relate to electrochemistry and are not considered in this cour

In conclusion, let us consider the structure of certain electrodes.

To read electrocardiograms, electrodes in the form of metal plates with terminals *1* are attached to the limbs with special rubber bands (fig. 21.3) into which pins of cables of leads are inserted and fixed. The cables connec the electrodes with the electrocardiograph. Electrode *2* is placed on the patient chest. A rubber suction cup holds the electrode. It also has a terminal for the lead cable pin.

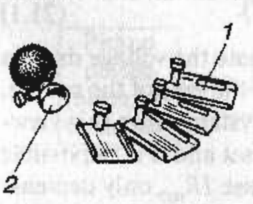


Fig. 21.3

In microelectrode practice, glass microelectrod are used. The contour of such an electrode is shown in fig. 21.4, its tip is 0.5 μm in diameter. The electrod frame is an insulator; there is a conductor in the for an electrolyte inside. Manufacturing microelectro and working with them is quite difficult; nevertheless, such a microelectrode makes it possible to pierce the cell membrane and do intracellular studies.

Fig. 21.4

§ 21.3. SENSORS OF BIOMEDICAL INFORMATION

Electrodes cannot read many biomedical characteristics, because they are not in the form of a bioelectrical signal: blood pressure, temperature, heart sounds, and many other parameters. Though in certain cases, biomedical information is connected with an electrical signal, however, it is more convenient to approach it as a non-electric quantity (for example, pulse rate). In such cases, sensors (measurement transducers) are used.

A sensor is a device that converts a quantity being measured or controlled into a signal that is convenient for transmission, further conversion, or recording. The sensor that measures a quantity directly, that is the first one in the measurement circuit, is called *primary* one.

In medical electronics, only sensors, which convert non-electrical quantity being measured or controlled into an electrical signal, are considered.

Use of an electrical signal is preferable to other signals because electronic devices make it possible to amplify, transmit and record signals quite easily. Sensors are divided into ***generating*** and ***parametric*** ones.

Generating sensors are those that under the influence of the quantity being measured, generate current or voltage directly. Let us enumerate some types of such sensors and phenomena, on which they are based:

1) *piezoelectric* ones, using piezoelectric effect (see Chapter 14);
2) *thermoelectric* ones, thermoelectricity (see Chapter 15);
3) *inductive* ones, electromagnetic induction (see Chapter 17);
4) *photoelectric* ones, photoelectricity (see § 27.8).

Parametric sensors are sensors, in which a parameter changes under the influence of the signal being measured. Let us enumerate some types of such sensors and the parameter measured with the sensors:

1) *capacitive*, capacity;
2) *rheostatic*, ohmic resistance;
3) *inductive*, inductance or mutual inductance.

Depending on the energy carrying information, there are mechanical, acoustic (sonic), thermal, electrical, optical, and other sensors.

In some cases, sensors are designated according to the quantity being measured; for example, pressure sensor, strain sensor for measurement of displacement or deformation, etc.

Let us represent possible biomedical application of these types of senso (Table 21.1).

Table 21.1

Sensor	Mechanical	Acoustic	Optical	Thermal
Piezoelectric	ABP	PCG	–	–
Thermoelectric	–	–	–	T
Inductive	BCG	PCG	–	–
Photoelectric	–	–	OHG	–
Capacitive	PCG	–	–	–
Rheostatic	AP, BCG	–	–	T
Inductive	PGT	–	–	–

Note. ABP — arterial blood pressure; BCG — ballistocardiogram; PCG — phonocardiogram; OHG — oxyhemography; T — temperature; PGT — pressure in the gastrointestinal tract.

A sensor is characterized by a ***conversion function*** — the functional depe dence of the output quantity y on the input one x, which is expressed by an analytical expression $y = f(x)$ or by a plot curve. The simplest and mos demonstrative case is a linear dependence $y = kx$.

Sensor sensitivity shows, to what extent the output quantity reacts to a hange in input value:

$$z = \Delta y / \Delta x.$$

Depending on the sensor type, the quantity is expressed in *Ohms per millimeter* (Ohm/mm), in *millivolts per Kelvin* (mV/K), etc. The sensitivity of a serial set of sensors is equal to the product of sensitivities of all sensors.

Time characteristics of sensors are essential. The point is that the physic processes in sensors do not occur instantly, so this leads to deceleration in the change of the output quantity compared with the change in the input quanti Analytically, this v leads to a dependence of the sensor sensitivity on the rate of change in the input signal dx/dt or on the frequency when x meets a harmonic law.

Dealing with sensors, one should take into account all possible errors which are inherent to them. Causes of errors can be as follow:

1) temperature dependence of the conversion function;
2) hysteresis, which is a lag of y behind x even at a slow change in the inp quantity, which occurs because of irreversible processes in the sensor;
3) variability of the conversion function over time;

4) feedback of the sensor on the biological system leading to a change in indications;
5) transit time of the sensor (neglect of its time characteristics), etc.

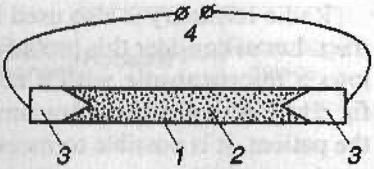


Fig. 21.5

The structure of sensors used in medicine is quite diverse: from simplest ones (like thermocouples) to complex Doppler sensors. As an example, let us describe a very simple respiratory rate sensor, which is rheostatic (resistive).

This sensor (fig. 21.5) is made in the form of a rubber tube *1*, which is filled with fine coal powder *2*. Electrodes *3* are installed at the ends of the tube. A current can be passed through the coal from an external source 4.

If the tube is put around the chest or, as it is usually done, a belt is attached to the ends of the tube and encircles the chest, then, during inhalation, the tube is stretched and during exhalation, it deflates. In this process, the electrical contact between particles of the coal powder changes, and the resistance of the sensor changes corresponding. The current intensity in the circuit will change too, which can be recorded using the corresponding measuring circuit.

In conclusion, we note that sensors are technical analogs of receptors of biological systems.

§ 21.4. SIGNAL TRANSMISSION. RADIO TELEMETRY

A read and amplified electrical signal should be transmitted to the recording (measuring) instrument.

In many cases, electrodes or sensors, amplifier and recording device are structurally designed as an integral device. In this case, transmission of information is not a technical problem. However, the measuring part can be located at a distance from the biological system. Such measurements are referred to as ***telemetry*** or even, in a narrow sense, as ***biotelemetry.*** In this case, communication between the reader and recording device is established either through wires or by radio. The latter option of telemetry is called ***radio telemetry.*** This type of communication is widely used in space exploration to obtain information about the actual state of a spacecraft and its crew, in sport medicine its is used to be informed about the physiological state of an exercising sportsperson. For example, using a transmitter antenna on the helmet of a sportsperson when the antenna emits radio waves, it is possible to record data about his or her actual state from a distance of 300–500 m (that is, within a stadium).

Radio telemetry is also used in endoscopic radio probing of the dig
tract. Let us consider this problem in more detail. The patient (fig. 21.7) swallows a microcapsule with a radio transmitter (endoscopic radio prob
fig. 21.6). Measuring the transmitter frequency with a receiver located near the patient, it is possible to measure the pressure, acidity or alkalinity, temperature and other parameters at the location of the capsule.

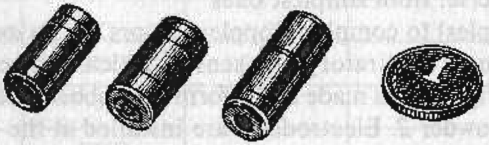


Fig. 21.6

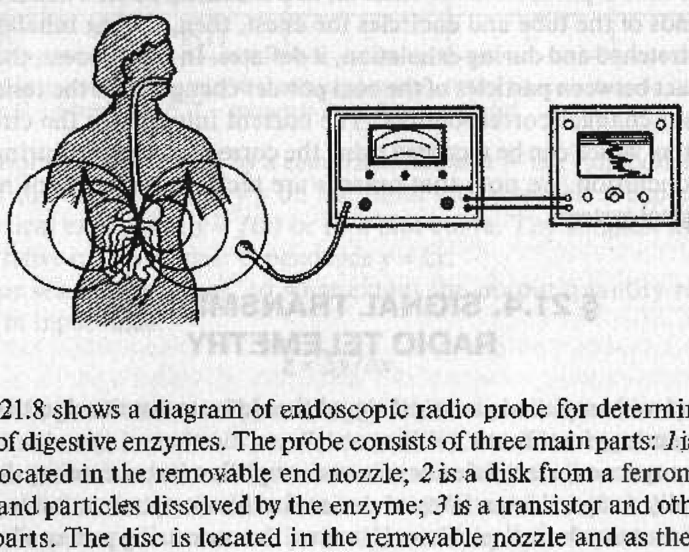


Fig. 21.7

Fig. 21.8 shows a diagram of endoscopic radio probe for determining the activity of digestive enzymes. The probe consists of three main parts: *1* is voltag
source located in the removable end nozzle; *2* is a disk from a ferromagne
powder and particles dissolved by the enzyme; *3* is a transistor and other radio circuit parts. The disc is located in the removable nozzle and as the voltag
source, it can be replaced with another after a single use. The principal elec
cal circuit of the generator is shown in fig. 23.1.

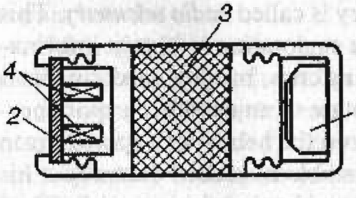


Fig. 21.8

The disc is pressed against the induc
tor *4* and forms a close-circuit magn
core together with it. As the disc is dissolved by digestive enzymes, inductance *L* of the circuit decreases and [see (18.7)] the frequency of the generator increas
Thus, the activity of enzymes can be estimated by the frequency received.

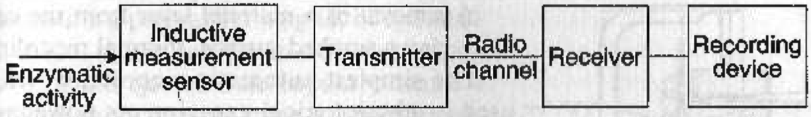


Fig. 21.9

The block diagram of an endoscopic radio probe is shown in fig. 21.9.

In some cases, multichannel devices are used. They permit receiving, transmitting, and recording several parameters simultaneously. Versatile biomedical information expands diagnostic capabilities very much. Nowadays, processing of such information is carried out with the help of computers.

§ 21.5. ANALOG RECORDERS

The terminal element of the technical diagram shown in fig. 21.1 is a measuring (monitoring) device that displays or records biomedical information.

A ***displaying device*** is a device that presents information for a short time; then when new information appears, the old information disappears without a trace. These are, in particular, needle instruments: ammeters, voltmeters, etc. A needle ammeter, for example, shows current intensity at the definite point of time and does not record it. When the current intensity in a circuit changes, information about the previous value disappears irrevocably. To memorize the information displayed by such devices, it is necessary to record it intentionally, which, for example, students do in the physical laboratory, taking readings of instruments. Biomedical applications of display devices are quite limited: resistance electric thermometer, pulse frequency meter, etc.

In medical electronics, much more common are ***recorders*** that write information onto a certain carrier. This permits documenting, storage, multiple use, processing and analyzing biomedical information obtained.

Displaying and recording devices are divided into ***analogous*** (continuous), ***discrete*** and ***combined*** ones combining the possibilities of analog and discrete devices.

Let us consider the most common analog recorders used in the practice of biomedical research in more detail. Some of them are also called self-recording instruments or automatic recorders.

In medicine, biology, and physiology, the following methods of recording information on a carrier are mainly used:

a) applying a layer of material (dye): ink-pen and jet writing systems;

b) changes in the state of the carrier material: photo recording, electrochemical, electrophotographic (xerographic) and magnetic recording;

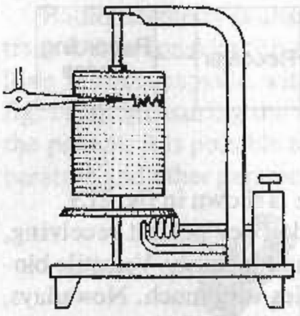


Fig. 21.10

c) removal of a material layer from the carrier: a smoked surface, thermal recordin

The simplest automatic recorder, which i used in physiological experiments nowaday is *kymograph* (fig. 21.10) working from a wound-up spring, or *electrical kymograph* with a dru which is rotated uniformly by an electric moto

The idea of kymograph — uniform rotatio or motion of a carrier surface — persists in mos modern analog recording devices, recordi the time dependence of a quantity under stud

Travel y of automatic pen or light spot is proportional to the quantity being recorded and is the ordinate of the curve obtaine (fig. 21.11). Constant speed of the carrier (paper, photo film) means that th abscissa is directly proportional to time t. As a result, the obtained curve reflects dependence $y = f(t)$.

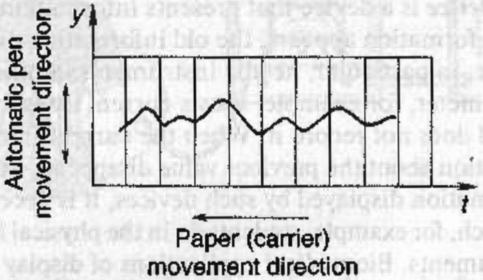

y
Automatic pen movement direction
t
Paper (carrier) movement direction

Fig. 21.11

Recorders used in medical equipment convert an electrical signal into mechanical motion. Physically, they are galvanometers — highly sensitive electric measuring instruments, which response to a quite low current intensit In these instruments, the current flowing through coils, filament frame or loop interacts with the magnetic field of a permanent magnet. Because of this interaction, the moving part (magnet, filament frame or a part of the loop) deflects proportionally to the current intensity, i.e. proportionally to the electrical signal.

A pen is connected to the moving part, the former leaves a trace on the moving recording carrier: special capillary pen, glass capillary with a nozzl in an ink-jet recorder or a mirror that reflects a ray or something else.

As an example, fig. 21.12 sketchily shows an ink-jet recorder. In this, *1* is the electromagnet through whose windings an electrical signal bein passes; *2* is a permanent magnet of cylindrical shape, rigidly connected with glass capillary *3*. Ink is ejected from capillary nozzle *4* under pressure, leaving trace *y*, which is proportional to the deflection of the permanent magnet and, therefore, to the current intensity in the electromagnet.

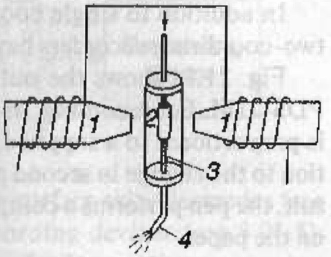


Fig. 21.12

An important characteristic of a recorder is the frequency range of oscillations that the device can record. The higher the angular momentum of the moving part of the recorder, the longer the lag of registration relative to the true change in the value recorded, and hence, the frequency response will be worse.

Analog recorders called *light-beam* (optical) oscillographs can perceive the widest frequency range.

The main part of the light-beam oscillograph (fig. 21.13) is loop galvanometer *G* consisting of permanent magnet *1*, metallic thread of loop shape *2*, along which the recorded electric signal flows, and small mirrors *3*. The ray from light source *Ls* falls onto a mirror and is reflected and falls onto photo device *Ph*, consisting of a roll of photo film and a film feed. The electrical signal causes a proportional deviation of the galvanometer loop. Uniform feeding film creates a time base. Using a special time markings can be obtained on the film.

Such an oscilloscope makes it possible to record phenomena with frequency range from about 0 to 10 kHz, which is much wider than frequency ranges of a biomedical signal. Mainly, multichannel light-beam oscillographs are produced; they make it possible to recording dozens of changing values simultaneously. Their main disadvantage is a need to develop photographic paper or photo films. At present, special paper, which is sensitive to ultraviolet rays, is produced. It does not require special treatment, but the light source should emit a powerful pencil of ultraviolet rays.

In automatic recorders, along with usual errors of measuring instruments, there are also errors due to recording itself.

Recording errors can occurs because of the following: inaccuracy of the paper or film feed, lag caused by the inertia of the instrument's recording system, change in paper dimensions upon exposure to air humidity, inaccuracy of the time marks, etc.

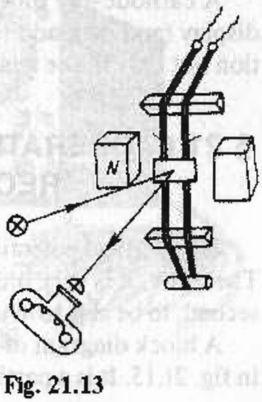


Fig. 21.13

In addition to single coordinate recorders registering time dependen two-coordinate recorders have become common in research practice.

Fig. 21.14 shows the outward appearance of such a recorder of mod PDS-21M. For recording, its cross rod moves translationally, its displacem is proportional to a supplied signal (parameter) x. Along the rod, in prop tion to the change in second parameter y, a carriage with a pen moves. As a result, the pen performs a complex motion and leaves a curve of function $y =$ on the paper.

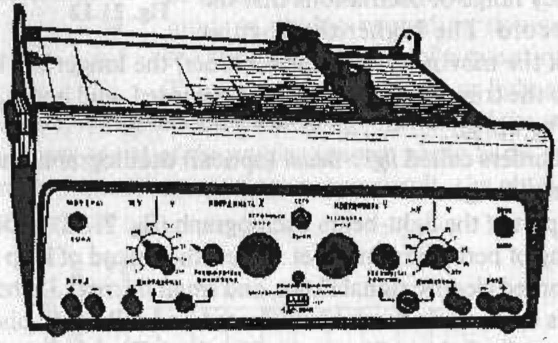


Fig. 21.14

Along with analog recording devices, such non-inertial combined devic as *cathode-ray tubes* are used in medical practice to record informati (see § 23.4).

For example, in a portable *vector-cardioscope* (see fig. 21.17), the cathod ray tube is the main element that displays, and with additional photographi records electro- and vector-cardiograms.

A cathode-ray tube belongs to the group of combined devices, as it can display (and with additional photographing even record) the output information not only in the analog form, but also in discrete one (digits, letters).

§ 21.6. OPERATION PRINCIPLE OF MEDICAL DEVICES RECORDING BIOPOTENTIALS

Bioelectrical potentials are an essential diagnostic indicator of many disease Therefore, it is very important, first, to record these potentials correctly and, second, to be able to extract necessary medical information from measuremen

A block diagram of medical instruments recording biopotentials is shown in fig. 21.15. It is a particular case of the general diagram shown in fig. 21.1.

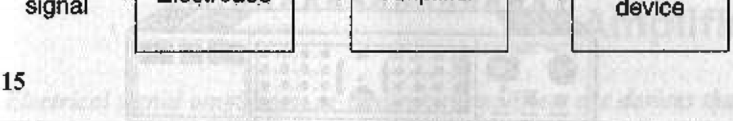


Fig. 21.15

In clinical practice, biopotentials are read by surface skin electrodes (see § 21.2), recording is carried out by analog recording devices (see § 21.5). A special switch makes it possible to pass from one lead to another.

Since biopotentials time relatively slowly over time, DC amplifiers are usually used in the devices (see § 22.5).

Biopotentials used in electrocardiography have values of about several millivolts; in electroencephalography, they can be about several microvolts, so for their recording, it is necessary to amplify them several thousand times, which is done by multistage amplification.

Fig. 21.16 shows the outward appearance of a portable transistor electrocardiograph EK-873 intended to record electrocardiograms, and fig. 21.17 shows a movable vector cardioscope VEKS-1p. This instrument makes it possible to study cardiac electrical activity by both electrocardiography and vector cardiography. The process can be visible on the screen of the cathode-ray tube and be photographed too.

In certain cases, it is expedient to determine a number of parameters simultaneously with one instrument, for example, biopotentials read from different points of the brain. In this case, multi-channel devices consisting of several independent amplifiers are used. Recording through all channels is made onto common film. Fig. 21.18 shows the outward appearance of a 16-channel encephalograph EEG16-01.

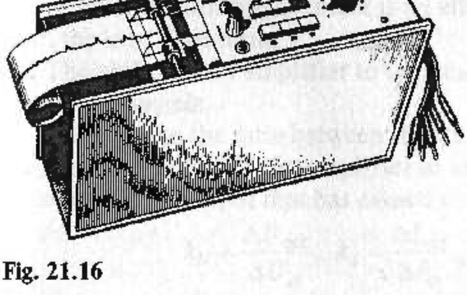


Fig. 21.16

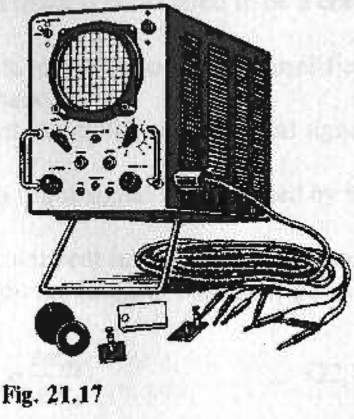


Fig. 21.17

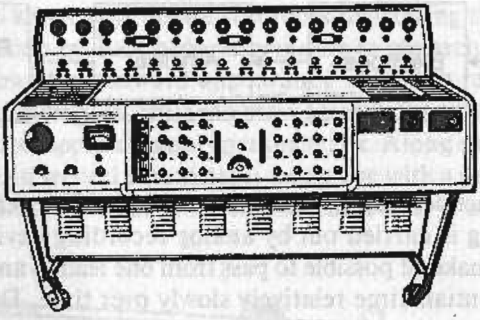


Fig. 21.18

For reading and recording biopotentials, certain auxiliary devices are used; they are absent in the block diagram in fig. 21.15. They include time markers that determine the scale of t axis (see fig. 21.11). In cases when the film feed ensures strict uniformity of the carrier motion speed, there is no need for a time marker.

To determine biopotentials, in other words, to determine the scale of axis y (see fig. 21.11) in voltage units, laboratory voltage calibrators are used. The calibration voltage is recorded before or after recording the biopotential. When taking an electrocardiogram a calibration signal of 1 mV is used.

Chapter 22

Amplifiers

Electrical signal amplifiers, or electronic amplifiers are devices that amplify these signals at the expense of the energy of an external source. Certain general problems relating to amplification of electric signals are discussed in this chapter, certain specific diagrams are described, and specifics of bioelectrical signals amplification is analyzed.

§ 22.1. AMPLIFIER GAIN

Amplifiers can be designed based on various elements (transistors, triodes, etc.), but, on the whole, all amplifiers can be represented quite uniformly. They have the input, to which the amplified electrical signal amplified, and an output, from which the amplified signal is read (fig. 22.1). A compulsory part of the entire system is a source of electrical energy.

The most common principle of signal amplification is action of the input circuit on the electrical resistance of the output circuit. This action corresponds to the shape of the signal being amplified, and therefore, the shape of the signal curve is reproduced in the output circuit.

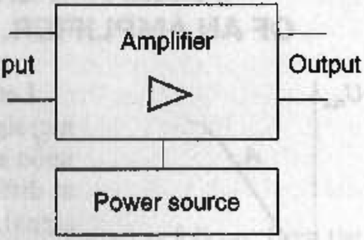


Fig. 22.1

In accordance with the concepts of cybernetics, the input circuit is considered to be a control system and the output circuit is considered to be a control object.

An essential requirement for amplifiers is reproduction of the amplified signal (gain) without distortion of its curve shape.

In practice, this requirement is an attempt to amplify the electrical signal with the least distortion.

The ability of an amplifier to increase an input signal is quantified by its ***amplification gain***.

It is equal to the ratio between the voltage (current intensity, power) increment at the output of the amplifier to the voltage (current intensity, power) increment at the input that has caused it:

$$k_U = \frac{\Delta U_{\text{out}}}{\Delta U_{\text{in}}}, \; k_I = \frac{\Delta I_{\text{out}}}{\Delta I_{\text{in}}}, \; k_P = \frac{\Delta P_{\text{out}}}{\Delta P_{\text{in}}}. \quad (22.1)$$

Depending on purpose, amplifiers are distinguished by their voltage, current intensity, or power.

Hereafter, for the sake of clarity, all illustrations and conclusions will refe to the voltage amplification gain, which will be designated without a subscrip

When a sine signal is amplified, in expressions (22.1) the amplitudes of the input and output signals are usually used:

$$k_U = U_{\max\,\mathrm{out}}/U_{\max\,\mathrm{in}}. \qquad (22.2)$$

If k is not sufficient to obtain an output signal of voltage required then several amplifiers are connected. In this case, each individual amplifier is calle an ***amplifier stage***. Essentially, fig. 22.1 and expressions (22.1) and (22.2 should refer to a stage. The gain of an amplifier with several stages is equa to the product of amplification gains of all stages used:

$$k_{\mathrm{tot}} = k_1 k_2 k_3 \ldots \qquad (22.3)$$

§ 22.2. AMPLITUDE CHARACTERISTIC OF AN AMPLIFIER. NONLINEAR DISTORTIONS

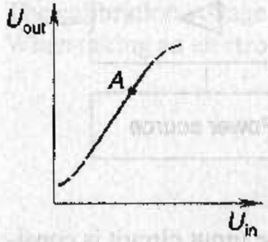


Fig. 22.2

Let us consider amplification of a sine (harmonic) signal. For the signal curve shape not to chan upon amplification, amplifier gain should be equa at different voltages within variation of the inp signal. In this case, dependence $U_{\mathrm{out}} = f(U_{\mathrm{in}})$ called an ***amplifier amplitude characteristic*** and is linear: $U_{\mathrm{out}} = kU_{\mathrm{in}}$ (fig. 22.2; straight line). In fact, a linear dependence is valid in a limited area of the input voltage range; going beyond this domai we have nonlinearity of the dependence (dashed line).

If the input harmonic signal goes beyond the linear part of an amplitud characteristic, the output signal will no longer be harmonic. Nonline (amplitude) distortions will occur.

Graphically, amplification of a harmonic signal is illustrated in fig. 22.3 without distortion (a) and with distortion (b).

Each of these figures shows three plot curves. On one (axes U_{in} and U_{out}), an amplitude characteristic is shown: linear (a) and non-linear (b). On the lower plot curve, the input voltage depends on time. This dependence is sinusoidal, but shifted relative to $U_{\mathrm{in}} = 0$ by a certain constant value. The curve is placed in unusual way, because there is common axis U_{in} with previous dependence. The left curve shows the time dependence of the output voltage. In this,

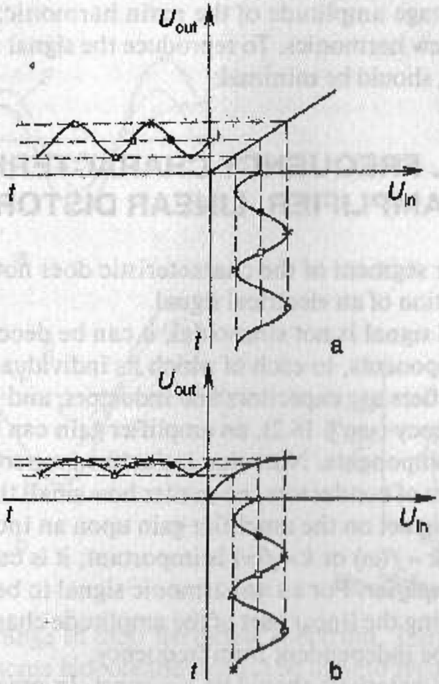


Fig. 22.3

axis U_{out} belongs to two plot curves. This curve is plotted as follows: from the lower plot curve, for certain selected points of time, values of U_{in} are found; after that, according to their amplitude characteristic, the corresponding values of U_{out} are calculated and they are translated on the left curve (dashed lines; points with the same symbols correspond to the same time).

On curves of dependence $U_{out} = f(t)$, in case of a linear amplitude characteristic (a), we can see a sine wave; hence, the signal amplified is not distorted. When the characteristic is nonlinear (b), the output signal is periodic, but is not sinusoidal; therefore, when the signal is amplified, it is distorted.

A periodic output signal obtained after nonlinear distortions of an input harmonic signal can be represented by a sum of harmonics (see § 7.4), so nonlinear distortions can be considered new harmonics arising in the signal upon its amplification. The more new harmonics, the higher their amplitude is, the stronger the nonlinear distortions are, which are calculated by a ***coefficient of nonlinear distortion***:

$$\gamma = \sqrt{U_2^2 + U_3^2 + \ldots} / U_1, \tag{22.4}$$

where U_1 is the voltage amplitude of the main harmonic; U_2, U_3, ... are th amplitudes of the new harmonics. To reproduce the signal accurately, this coefficient, obviously, should be minimal.

§ 22.3. FREQUENCY CHARACTERISTICS OF AN AMPLIFIER. LINEAR DISTORTIONS

Using the linear segment of the characteristic does not yet guarantee undistorted amplification of an electrical signal.

If the amplified signal is not sinusoidal, it can be decomposed into separate harmonic components, to each of which its individual frequency corresponds. Since amplifiers use capacitors and inductors, and their resistance depends on the frequency (see § 18.2), an amplifier gain can be different for different harmonic components. Note that inductive properties of resistors and capacitive properties of conductors, no matter how small they can be, can also exert an essential impact on the amplifier gain upon an increase in frequen Thus, dependence $k = f(\omega)$ or $k = f(\nu)$ is important; it is called ***frequency characteristics of an amplifier***. For an anharmonic signal to be amplified witho distortions (even using the linear part of the amplitude characteristic), the amplifier gain should be independent from frequency.

A frequency characteristic should be k = const. In practice, this is impossible and leads to distortions called **linear** or **frequency** ones.

Linear distortions are shown in fig. 22.4. Fig. 22.4, a, shows periodic signal *3*, which is a sum of two sine waves (*1* and *2*). If sinusoidal signals are amplified differently, for example, one with $k_1 = 2$ and the other with $k_2 = 0.5$, then the resulting amplified signal differs from the input one (compa curves *3* in fig. 22.4, a, b).

The frequency characteristic of an amplifier is usually depicted graphical (fig. 22.5). In this figure, we can see that between ω_2 and ω_3, the amplifier ga is approximately constant. In radio engineering, it is accepted that reducin it to $0.7k_{\max}$ (or $k_{\max}/\sqrt{2}$) does not practically distort a signal. Frequency rang $\omega_1 - \omega_4$ is called ***amplifier bandwidth***.

To expand the bandwidth it is necessary to complicate amplifier circuit However, the frequency bandwidth to be amplified without distortions is determined by the amplification purposes. For instance, 60 Hz–15 kH bandwidth is sufficient for sound amplification but video signals require amplification of a much wider bandwidth.

The frequency characteristic is of great importance in choosing an amplifi for recording biopotentials, whose oscillations are very complicated with dif-

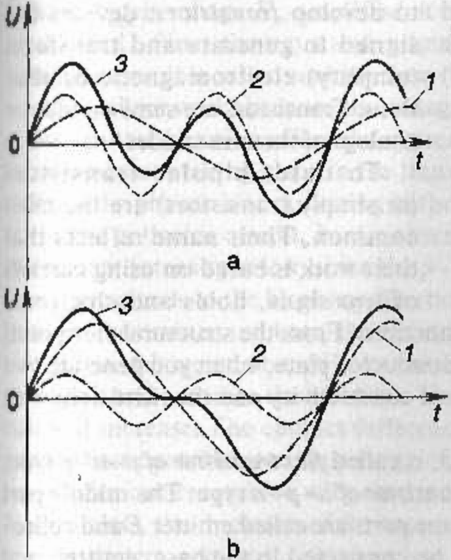


Fig. 22.4

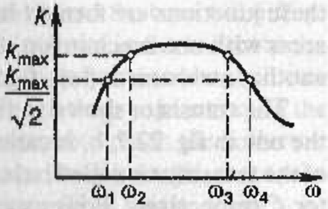


Fig. 22.5

ferent frequencies range in their harmonic spectrum. Therefore, an amplifier intended to record some biopotentials could be used to record some other type.

§ 22.4. TRANSISTOR AMPLIFIER

Let us consider a transistor as a specific physical device that amplifies an electrical signal.

At the interface of two semiconductors with different types of conductivity, *p–n*-junction occurs. This is a small area in thickness, the resistance of which depends on the direction of applied voltage. A current-voltage characteristic of the *p–n* junction is shown in fig. 22.6, there is a conventional sign of a semiconductor diode there. The forward voltage (positive voltage on the curve) is matched by the forward direction of the current; the reverse voltage (negative on the curve) is matched by the blocking direction of the current.

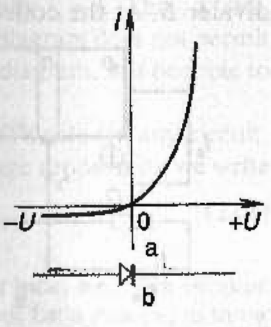


Fig. 22.6

Contacts of semiconductors with different types of electrical conductivity can also be used

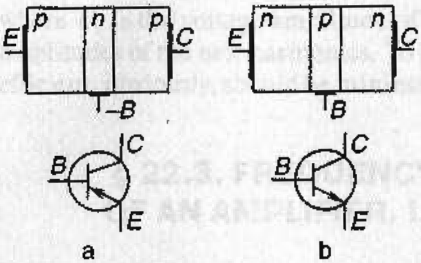


Fig. 22.7

to develop *transistors*: devices designed to generate and transfo (amplify) electromagnetic oscill tions. Transistors are semi-condu analogs of vacuum triodes.

To date, bipolar transisto (or simply transistors) are the mo common. Their name reflects tha their work is based on using carri of two signs, holes and electro These transistors consist of two p–n junctions. From the structural viewpoi these junctions are formed in a semiconductor plate, when you generate tw areas with one mechanism of electrical conductivity and the third area wit another mechanism (top, fig. 22.7).

The transistor shown in fig. 22.7, a, is called ***flat transistor of p–n–p*** typ the one in fig. 22.7, b, is called ***flat transistor of n–p–n*** type. The middle par of the transistor is called base B, extreme parts are called emitter E and colle tor C respectively. Power supply can be connected to the base, emitter, an collector by means of metal electrodes. At the bottom of fig. 22.7, conve tional signs of the corresponding electric circuit transistors are show At the emitter, the arrow points to the direction of the current; the *p–n junction* between the emitter and base is called emitter junction, and the arrow between the collector and base is called collector junction.

Let us consider the physical operational principles of a transistor and its ability to amplify an electrical signal with an example of a p–n–p-transis and the circuit shown in fig. 22.8 (a — schematic representation of transis b is the conventional representation of transistor).

Current source E_e is connected to emitter circuit I_e, it generates direc voltage on the emitter junction. This voltage U_{in} can be changed using volta divider D. At the collector junction, source E_c generates reverse voltag

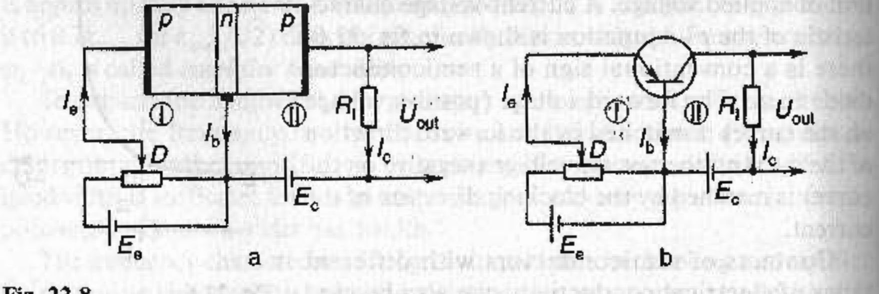


Fig. 22.8

Voltage U_{out} is voltage U_{in} amplified; it is read from load resistor R_b in collector circuit II. Since the input and output voltages have a common point at the base, this circuit is called ***common base*** (CB).

A transistor works because the voltage and, therefore, the current intensity at the emitter junction affect the current in the collector circuit. To understand processes taking place in the transistor it is necessary to recall the specifics of the phenomena occurring in the p–n-junction.

In the collector circuit, in the absence of voltage, in the emitter circuit, the current intensity is low, since the p–n-junction between the collector and base corresponds to the blocking direction. If we generate and increase voltage between the emitter and base, the current intensity will increase in the emitter circuit too. A large number of holes passing (injected) into the base will diffuse through it and pass into the collector. The current intensity in the collector circuit will increase. The contact difference of potentials between the collector and the base contributes to the passage of holes through the p–n junction. In the base, holes can recombine with electrons and cannot but reach the collector. Nevertheless, the thickness of the base is made to be small enough, about tens of micrometers, and most holes pass into the collector. Thus, in the emitter circuit, the current intensity affects the resistance of the collector junction.

The word ***emitter*** reflects the fact that the main current carriers (in this example, they are holes) are as if emitted from this electrode into the base. In fact, there is no emission, but injection of holes[1]. The word *collector* refers to the domain whose function is to extract charge carriers from the base. The base, according to the valid terminology, is a domain, into which the emitter injects charge carriers that are minor for this area.

At different segments of the circuit, the current directions are shown in fig. 22.8; the ratio between the current intensities is as follows:

$$I_e = I_c + I_b. \tag{22.5}$$

Since the base current intensity is low then practically $I_e \approx I_c$. Thus, a transistor connected into the circuit according to the CB diagram does not permit amplifying the current intensity. However, using this diagram, it is possible to amplify the voltage and power. Let us explain it.

For circuit II of the collector, based on Kirchhoff's rule (in any circuit, the sum of emfs of sources is equal to the sum of voltage drops in it), we write

$$\mathcal{E}_c = I_b R_b + I_c R_c + I_c R_1, \tag{22.6}$$

---

[1] In physics, the term "emission" (from Latin *emission* for issue) means for emission of electrons from metals into vacuum or gas. Injection (from Latin *injacere*, to throw inside) means passage of current carriers through a p–n-junction.

where R_1 is the resistance of the circuit segment with current intensity I_b of th base; R_c is the resistance of the collector junction against current I_c. We do no take into account the voltage drop at the source.

Since the base current intensity is low, [see (22.6)]:

$$\mathcal{E}_c = I_cR_c + I_cR_1. \qquad (22.7)$$

This means that the emf of the collector circuit source is equal to the su of the voltage drops at the collector p–n junction and load resistor. It is eas to calculate the dependence of the output voltage on parameters of the colle tor circuit. From (22.7), we have

$$I_c = \mathcal{E}_c/(R_1+R_c).$$

Substituting this expression into the formula for an output voltage, w obtain:

$$U_{\text{out}} = I_cR_1 = \frac{\mathcal{E}_cR_1}{R_1+R_c} = \frac{\mathcal{E}_c}{1+R_1/R_c}. \qquad (22.$$

As we can see from (22.8), variations of U_{out} are possible in the limits fro zero at $R_c \gg R_1$ to E_c at $R_c \ll R_1$. As it previously noted, a change in R_c depends on a change in the current in the emitter circuit and, therefore, on th change in input voltage. Therefore, we obtain dependence $\Delta U_{\text{out}} = f(\Delta U_i$ which at certain conditions, can be directly proportional:

$$k = \frac{\Delta U_{\text{out}}}{\Delta U_{\text{in}}}, \qquad (22.9)$$

where k is the amplifier voltage gain.

Amplification capabilities of a transistor depend not only on its properti and structure elements (material, impurity concentration, base dimensio etc.), but also on the elements of its amplifying stage. However, it is possib to single out parameters that depend only on the transistor itself. One of thes parameters is, in particular, ***current gain*** for the CB circuit (current transmi sion coefficient of the emitter), which depends on the thickness of the ba and parameters of the semiconductor material. This coefficient is equal the ratio between an increment of the collector current intensity to the chan in the emitter current intensity that caused it while the collector–ba voltage remains constant:

$$\alpha = \frac{\Delta I_c}{\Delta I_e} \quad \text{at } U_{\text{cb}} = \text{const}.$$

Since $I_c \approx I_e$, $\alpha \approx 1$. At a low frequency the change in input voltage α is 0.95–0.998. A constant voltage on the collector p–n-junction can be generat short-circuiting the load resistor. As we can see from (22.7), in this case $U_{cb} \approx \mathcal{E}_c$, if the internal resistance of the source is low.

The input (amplified) power is always supplied to the emitter-base couple, but this does not mean that through the emitter and the base, input current flows, since either the emitter or the base is connected to the collector circuit. The current intensity of an input electric signal is equal to I_e when the only input conductor is connected to the emitter. This variant corresponds to a CB circuit.

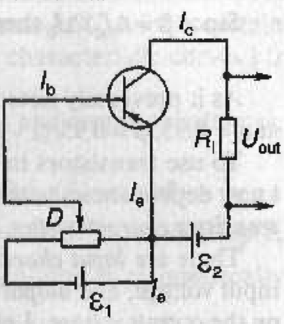


Fig. 22.9

There is an option of circuits, in which only one conductor of input signal is connected to the base. In this case, the current intensity of an input electrical signal is equal to I_b. Fig. 22.9 shows such a circuit called ***common emitter*** (CE) circuit. In this circuit, the emitter is the common point of input and output. Physical operational principles of the transistor remain the same in this circuit too: the emitter current affects the resistance of the collector junction.

In such a diagram, amplifying properties of the transistor itself are characterized by ***current gain*** for the CE circuit (current transmission coefficient of the base). It is equal to the increment of current intensity ΔI_c of the collector that caused the current intensity increment ΔI_b of the base at a constant collector-emitter voltage:

$$\beta = \frac{\Delta I_c}{\Delta I_b}$$

(U_{ce} = const). A constant voltage can be generated, for example, shortening a load resistor ($R_l = 0$). In CE circuit, ΔI_b corresponds to the change in input current, and ΔI_c corresponds to the change in output current.

Coefficients α and β depend on the transistor properties, and reflect the connection among the currents of the emitter, collector, and base. Let us determine the relation between these coefficients.

From (22.5), it follows that $\Delta I_e = \Delta I_c + \Delta I_b$. Substituting this expression into the formula for α:

$$\alpha = \frac{\Delta I_c}{\Delta I_e} = \frac{\Delta I_c}{\Delta I_e + \Delta I_b}.$$

Dividing the numerator and the denominator by ΔI_b, we obtain:

$$\alpha = \frac{\Delta I_c / \Delta I_b}{\Delta I_c / \Delta I_b + 1}.$$

Since $\beta = \Delta I_c / \Delta I_b$ then

$$\alpha = \beta/(\beta+1), \quad \beta = \alpha/(1-\alpha).$$

As it previously noted, α is close to 1; therefore, $\beta \gg 1$. So, for examp at $\alpha = 0.95$, $\beta = 0.95/(1 - 0.95) \approx 19$.

To use transistors in specific amplifier or generator circuits we nee know dependencies between currents and voltages, which are generally ca ***transistor characteristics.***

There are ***input characteristics*** — dependences of the input current on th input voltage, and ***output characteristics*** — dependences of the output curre on the output voltage. Let us consider static characteristics of a transistor c nected according to a CB circuit. Such characteristics are read using an installation similar to one in fig. 22.8; one should only foresee adjusting he voltage not only at the emitter junction, but also at the collector junctio to the extent of changing their polarity, as well as the presence of measuring instruments — voltmeters and ammeters. The characteristics are determin at the constant current without a load in the output circuit ($R_l = 0$).

The output current in CB circuit is current intensity I_e of emitter, the inpu voltage is voltage U_{eb} between the emitter and base. That is why, the input characteristic is dependence $I_e = f(U_{eb})$ at U_{cb} = const, a family of input characteristic curves is shown in fig. 22.10. Essentially, these are current-voltage characteristics for the direct voltage of p–n-junction (see fig. 22.6; $U > 0$). Strict speaking, only one characteristic will pass through the origin of coordina (at $U_{cb} = 0$)ж however, these details are not shown in fig. 22.10. At differe values of U_{cb}, the characteristic curves are rather close one to another since th collector junction does not affect the emitter current intensity much.

For I_{ei}= const ($I_{e5} > \ldots > I_{e4} > \ldots > I_{e1}$), the family of output characteris curves $I_c = f(U_{cb})$ is shown in fig. 22.11. The lower curve ($I_e = 0$) is the curre voltage characteristic curve of the reverse voltage of the p–n-junction, (se fig. 22.6; $U < 0$). The output characteristic curves confirm what was previo

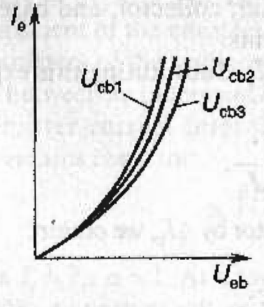


Fig. 22.10

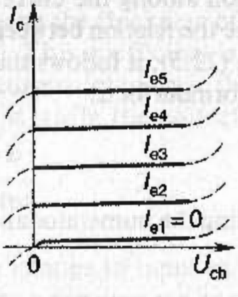


Fig. 22.11

noted: $I_c \approx I_e$, $\alpha \approx 1$; I_c changes poorly upon an increase in U_{cb} until a certain value, at which a spark-over starts (dashed lines on characteristic curves) is attained.

Essential parameters of a transistor are its input and output resistances, which are determined based on Ohm's law as follows:

$$R_{in} = \Delta U_{eb}/\Delta I_e \ (U_{cb} = \text{const}), \ \ R_{out} = \Delta U_{cb}/\Delta I_c \ (I_e = \text{const}).$$

These resistances can be expressed as derivatives; for example, $R_{in} = \dfrac{dU_{eb}}{dI_e} = \dfrac{1}{dI_e/dU_{eb}}$. Since a derivative of the function is geometrically equal to the slope of the tangent line to the curve relative to axis OX (in this example, relative to axis U_{eb} the input voltage can be written also as follows: $R_{in} = 1/\tan\varphi$, where φ is the angle of the tangent line to the curve (characteristic one) relative to axis U_{eb} (or $R_{in} = \operatorname{cotan}\varphi$). The last formula makes it possible to find the input resistance of a transistor graphically based on characteristic curves.

Similarly, it is possible to find output resistance based on the corresponding output characteristic curves.

For the CE diagram, static characteristic curves of the transistor are shown in fig. 22.12 (input) and 22.13 (output). The input characteristic curve is the dependence of current I_b of the base on voltage U_{be} (base-emitter): $I_b = f(U_{be})$ at $U_{ce} = \text{const}$ ($U_{ce1} > U_{ce2} > U_{ce3}$). The output characteristic curves are dependence $I_c = f(U_{ce})$ at a constant base current ($I_b = \text{const}$; $I_{b5} > I_{b4} > \ldots > I_{b1}$). Similar to the foregoing, based on the input characteristic curves, for the CE diagram, input resistance of the transistor can be found graphically, and on base of the output characteristic curve, we can find the output resistance and current transmission coefficient β of the base.

The outward appearance of a bipolar transistor is shown in fig. 22.14.

Further improvement of transistors led to development of field-effect transistors, which are nowadays represented by different variants. Let us consider

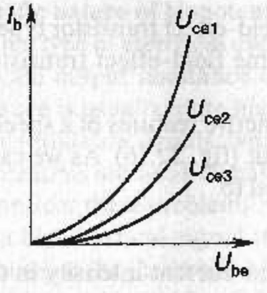


Fig. 22.12

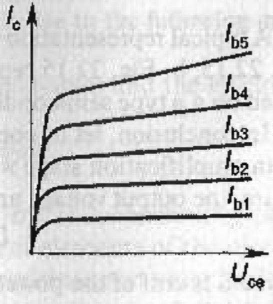


Fig. 22.13

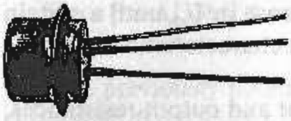


Fig. 22.14

a field-effect transistor with a controlling p–n-junction. For such a field-effect transistor, the basic principle of operation is the effect of the electr field (hence the attribute "field") of the input circuit on the width of the semiconductor chann through which the current of output circuit passe

To a semiconductor of n-type (the cross-section of the former is shown in fig. 22.15, a), source (S) and drain (D) electrodes are connected. To S-D circuit, constant voltage source ε and load resistor R_l are connected. In the bod of the main semiconductor, there is a semiconductor of p type, to which a lead, which is an electrode called gate G, is connected. At the interface of semiconductors with different conductance mechanisms, p–n-junction is formed: it is an area depleted of charge carriers and, therefore, carrying sizeable resistanc Since the width of a p–n-junction depends on the electric field, changin voltage U_{in} between the source and gate, it is possible to change the width of th channel, through which the current between the source and drain passe Thus, the input voltage influences the output one, which is read from resistor R_l. Field-effect transistors have a large input resistance, so in their case, upon amplification of biopotentials, problems of resistance matching are solved easier [see § 22.5].

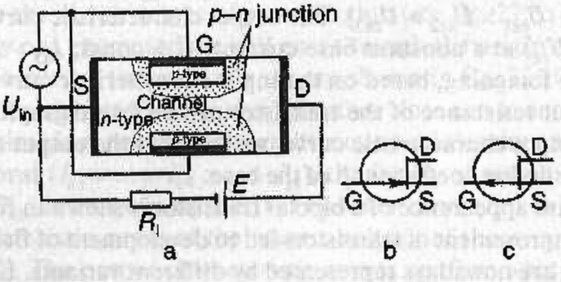


Fig. 22.15

A typical representation of the above field-effect transistor type is shown in fig. 22.15, b. Fig. 22.15 represents the same field-effect transistor designe based on a p type semiconductor.

In conclusion, let us consider the distinctive features of a specific diagra of an amplification stage with a CE circuit (fig. 22.16). As we can see in the figure, the output voltage amplified is equal to

$$U_{out} = \mathcal{E} - I_c R_c, \qquad \text{(22.1}$$

where Ɛ is emf of the power source; I_c is the current intensity in the collect circuit; R_c is the resistance of the resistor in the collector circuit.

A variable input signal is connected to the emitter junction and influences the collector current intensity, and, therefore, the actual voltage drop at the resistor of the collector circuit ($I_c R_c$). Thus, the output voltage [see (22.10)] changes according to the change in the input voltage. Coupling capacitor C_{s1} does not permit the constant component of the base current to flow through the input power source; capacitor C_{s2} filters only the alternating component to pass through the output.

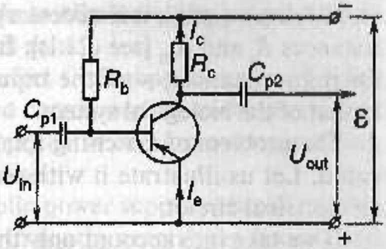


Fig. 22.16

To adjust a directly proportional dependence between the output and input voltages, on the amplitude characteristic curve (see fig. 22.2), it is necessary to find such a position of original operating point A so that the change of the input and output voltages are expressed by a linear function. When calculating the operating point coordinates, the characteristic curves, shown in fig. 22.12 and 22.13, are used. Resistor R_b of the base circuit (see fig. 22.10) makes it possible to set up a desired mode of operation of the circuit.

§ 22.5. AMPLIFICATION OF BIOELECTRICAL SIGNALS

Direct measurement of a bioelectrical signal or a signal generated at the output of a generating sensor is a difficult task, as these signals are usually very weak. That is why, in the block diagram (see fig. 21.1), the second element is an amplifier of electrical signals.

For the sake of clarity, let us illustrate some specifics of amplification of biomedical electrical signals with an example of ***bioelectrical signals***, i.e. signals that are read by electrodes from a biological object.

The specific nature of biopotential amplifiers is due to the following main features of this type of electrical oscillations.

1. The total output resistance of the biological system and the electrode resistance is usually quite high.
2. Biopotentials are signals, which alternate slowly.
3. Biopotentials are weak signals.

Let us consider these problems in more detail. In § 21.2, concerning transmission of a bioelectrical signal to the subsequent elements of the circuit, the importance of the electrode-skin transition resistance was emphasized. It turns out the problem of such resistances igoes well beyond that: when amplifying a signal, it is necessary to take into account the ratio between re-sistances R_i and R_{in} [see (21.1)]. It is necessary, as they say, to match the following resistances — of the input circuit, amplifier, and output circuit — to that of the biological system.

The problem of matching total resistances (impedances) is quite comp cated. Let us illustrate it with only two features of interacting elements of an electrical circuit.

If we take into account only the radio-engineering aspect of the proble then it is necessary to supply the maximum power to the input of an amplifi from the source of oscillations to be amplified. This is achieved when the internal resistance of the source is the same as the proper resistance of the con sumer, in our particular case, if $R_i = R_{in}$.

In other words, expressing the current intensity from (21.1) $I = \mathcal{E}/(R_i + R_{in})$ and substituting it into the formula for the amplifier input voltage, we obta

$$U = IR_{in} = \frac{\mathcal{E}_{bp}}{R_i + R_{in}} R_{in} = \frac{\mathcal{E}_{bp}}{R_i / R_{in} + 1}. \quad (22.11)$$

From (22.11), certain limiting cases follow:

1) $U_{in} \to 0$ when $R_{in} \to 0$, i.e. there will be no voltage at the amplifi input, if its input resistance is zero.
2) $U_{in} \to \mathcal{E}_{bp}$ when $R_{in} \to \infty$, i.e., at the amplifier, the maximum possib voltage will be when the input resistance is infinitely great.

In the second case, the input circuit will carry no current; therefo no power will be transmitted from the signal source.

In a real situation, the input of the amplifier will have a part of the powe generated by the biological system and depending on ratio R_i/R_{in} [see (22.11

The foregoing examples highlight a need to match resistances. In electr physiology, R_{in} is considered to be 10–20 times higher than the maxim possible value of R_i.

From the low frequency of biological electrical signals, it follows that it is impossible to use capacitors in the amplifier circuits, since at low frequencie the capacitive resistance increases to a great extent [see (18.28)]. We hav to use special DC amplifiers.

The low values of biopotentials necessitates the use of amplifiers with quit high gain, so even minor interferences, getting at the input of the amplifie can distort the useful bioelectrical signal and cause misinformation. Any cu rents or voltages at the output of the amplifier in the absence of a useful sig at its input are considered interferences.

It is possible to eliminate some interference or reduce it by making an ampli fier more sophisticated and increasing the cost of its raising its structur

Certain types of interference are impossible to eliminate altogether, so it is necessary to use a stage amplifier with a small gain.

For example, if a rectifier with an imperfect filter supplies an amplifying circuit, then a background noise is generated corresponding to the oscillation period of the current of the public power supply network. Improvement of rectification and surge control eliminate this noise.

The alternating current (AC) of the public power supply network can induce emf due to electromagnetic induction in nearby amplifying circuits and biological objects. Shielding the amplifier and the conductors in its circuit, as well as the systems being studied, removal of these elements from conductors with alternating current can eliminate or reduce this interference.

If the amplifier components (tube electrodes, capacitor plates, etc.) vibrate, this will lead to alterations of the parameters of the circuit and, as a result, to occurrence of random electromagnetic oscillations. This phenomenon is called ***microphonics***. By fixing the components of the circuit and heightening its shock absorption, it is possible reduce or eliminate interference of this type.

A large group of interference is called ***noises***. We can hear a noise, for example, in a radio set like hissing, crackling and rustling. This is especially noticeable at high amplification.

A noise is audible only when its frequency is 20 Hz–20 kHz. If we plot the dependence curve of the output current or voltage on time, then even in the absence of a signal at the input, the current and voltage will not be constant (fig. 22.17).

This is a manifestation of noises.

The term "noise" originated from auditory perception of electrical chaotic signals when a loudspeaker is connected. However, nowadays, this concept is used irrespective of its frequency range and even more irrespective of the fact whether the noises are reproduced by sound waves or not.

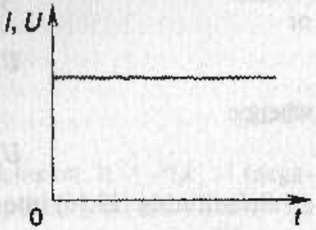


Fig. 22.17

Noises are caused by fluctuations of current intensity, i.e. random deviations from their average values; chaotic motion of electrons generates such deviations. It is quite difficult to get rid of noises; a lot of effort goes into manufacturing special low-noise transistors and tubes for use in the first stage of amplifier.

Distortions of the signal amplified can also occur due to instability of the power supply, so in certain cases, it is necessary to provide voltage stabilization specially. Let us consider the most common amplifiers used for biomedical tasks.

Feedback in electronic amplifiers. Common collectors. Amplifiers with very negative feedback are used for biomedical purposes.

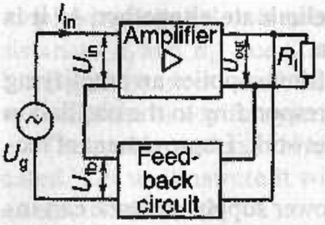


Fig. 22.18

Negative feedback impedes developme any changes in the process and stabilizes it. It is also important for electronic devices.

In the case of an amplifier, feedback mean the effect of a signal from its output on its input

A possible block diagram of the amplifi with feedback is shown in fig. 22.18. In this the feedback circuit is connected to the outpu of the amplifier in parallel to its load; henc voltage U_{fb}, of the feedback circuit is directly proportional to voltage U_{o} at the output. In the input circuit, the feedback voltage segment is connect in series with voltage U_g segment of the signal source (generator).

Let us calculate the gain of the amplifier with feedback[1]. Let us call the rati

$$\beta = U_{fb}/U_{out} \qquad (22.12)$$

transmission coefficient of feedback circuit, or $U_{fb} = \beta \cdot U_{out}$.

The amplifier gain of a circuit with feedback k_{fb} is equal to the rati between output voltage U_{out} to the voltage U_g of the signal source:

$$K_{fb} = U_{out}/U_g \qquad (22.1$$

Input voltage of the amplifier (fig. 22.18) is equal to

$$K_{fb} = U_{out}/U_g \qquad (22.14)$$

or

$$U_{in} = U_g + \beta U_{out}, \qquad (22.15)$$

whence

$$U_g = U_{in} - \beta U_{out}. \qquad (22.16)$$

Substituting (22.16) into (22.13), we obtain:

$$k_{fb} = U_{out}/(U_{in} - \beta \cdot U_{out}).$$

In this expression, dividing the numerator and denominator by U_{in}:

$$k = \frac{U_{out}/U_{in}}{1 - \beta\, U_{out}/U_{in}}. \qquad (22.17)$$

The ratio

$$k = U_{out}/U_{in}$$

---

[1] For the sake of clarity, phase relations between different voltages are not taken into account here.

is voltage ***gain*** of the amplifier proper, without feedback circuit. As a result, we have

$$k_{fb} = \frac{k}{|1-\beta k|}. \tag{22.18}$$

For the amplifier gain to be positive, in the denominator of formula (22.18), we use the absolute value of the differernce.

Let us analyze expression (22.18).

If $|1-\beta k| > 1$, then $k_{fb} < k$. The latter is valid when $\beta < 0$ since k is always positive. In this case, negative feedback is implemented, which leads to reduction of amplification capabilities of the circuit. Physically, this phenomenon occurs because the feedback voltage attenuates the voltage applied to the input of the amplifier.

When $|1-\beta k| < 1$, $k_{fb} > k$, $\beta > 0$, hence, the feedback is positive.

If $|1-\beta k| \to 0$, then $k_{fb} \to \infty$. From (22.13), it follows that when U_{out} is given, $U_g \to 0$, that is the voltage of electric signal source is equal to zero. Such a diagram can work without an external source of oscillations, that is, as a generator (rather than an amplifier!) with feedback.

All three consequences from (22.18) occur in practice, but medical electronics is especially interested in amplifiers with negative feedback.

Differentiating (22.18), we obtain:

$$\mathrm{d}k_{fb} = \frac{(1-\beta k)+\beta k}{(1-\beta k)^2}\mathrm{d}k = \frac{\mathrm{d}k}{(1-\beta k)^2}. \tag{22.19}$$

Dividing (22.19) by (22.18), we obtain:

$$\frac{\mathrm{d}k_{fb}}{k_{fb}} = \frac{1}{1-\beta k}\frac{\mathrm{d}k}{k}, \tag{22.20}$$

Expression (22.20) leads to the interesting conclusion: If $|1-\beta k| > 1$ (negative feedback), then the relative change in the amplifier gain of a circuit with feedback is $(1-\beta k)$ times less than the relative change in the gain of the amplifier proper:

$$\frac{\mathrm{d}k_{fb}}{k_{fb}} < \frac{\mathrm{d}k}{k}.$$

The above means higher stability of the amplifier with negative feedback, its lesser dependence on external influences. Let us illustrate this with a quantitative example. Let amplifier gain $k = 100$, transmission coefficient β of the feedback circuit is equal to –0,1. Substituting these values into (22.18), we have:

$$k_{fb} = \frac{100}{1-(-0.1)\cdot 100} = \frac{100}{11} \approx 9.1.$$

If for whatever reason, the gain of amplifier is reduced by 20% and becom equal to $k_1 = 80$, then the new value of the gain of the amplifier circuit wit feedback is

$$k_{\text{fb}1} = \frac{80}{1-(-0.1)\cdot 80} = \frac{80}{9} \approx 8.9.$$

k_{fb} is decreased only by

$$\frac{9.1-8.9}{9.1}100\% \approx 2.2\%(!),$$

although the amplifying capabilities of a circuit with negative feedback are les than the amplifying capabilities of the amplifier proper.

The input resistance of an amplifier with series feedback accordi to Ohm's law is equal to

$$R_{\text{in.fb}} = U_{\text{g}}/I_{\text{in}}.$$

Using (22.16), we obtain:

$$R_{\text{in.fb}} = \frac{U_{\text{in}} - \beta U_{\text{out}}}{I_{\text{in}}} = R_{\text{in}}(1-\beta k), \qquad (22.2[illegible])$$

where $R_{\text{in}} = U_{\text{in}}/I_{\text{in}}$ is the input resistance of the amplifier proper.

In case of negative feedback ($\beta < 0$), $R_{\text{in. fb}} > R_{\text{in}}$. With "deep", 100%, negative feedback ($\beta = -1$), we have $R_{\text{in. fb}} = R_{\text{in}}(1+k)$. At high values of k, $R_{\text{in. fb}} \gg R_{\text{in}}$.

Two specific features of negative feedback amplifiers (stability to ward external influences and high input resistance) make it possible to use such devices for purposes of medical electronics. Amplifiers with 100% negativ feedback are called ***common collectors***. They are used as intermediate amplifiers (preamplifiers) and are placed between a biological system and the mai amplifier for matching resistances. A high input resistance of common coll tors matches with the high resistance of the object being measured. The low output resistance of the common collector matches the low input resistan of the subsequent main amplifier. For this reason, common collectors are als called ***impedance transformers***.

From (22.18), it follows that for common collectors ($\beta = -1$)

$$k_{\text{fb}} = k/(1+k);$$

therefore, $k_{\text{fb}} \approx 1$ when $k \gg 1$. In practice, k_{fb} is approximately equal to 0.95–0.99. At such values, from (22.13), we obtain $U_{\text{out}} \approx U_{\text{g}}$; the output voltage equal to the voltage fed on the amplifier with 100% negative feedback. In th case, the phases of the input and output voltages are equal too. Hence the ter ***common collector***, more precisely — ***voltage follower***.

An implementation of a follower depends on the type of the electronic device used in it. There are the following three types of followers: ***source follower*** based on a field-effect transistor (fig. 22.19, a); ***emitter-follower*** based on a bipolar transistor (fig. 22.19, b); ***cathode follower*** based on an electron tube (fig. 22.19, c). The name of a follower depends on the name of the circuit (the circuit itself is named after the element of the electronic device), to which the load resistor R is connected (source, emitter, or cathode circuit).

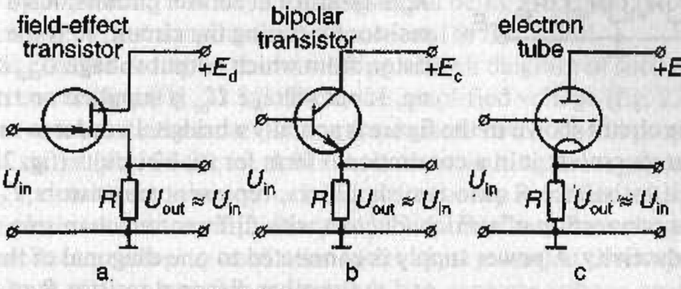


Fig. 22.19

In all these cases, you can make sure that the output voltage — the voltage on resistor R — is included completely (100%!) into the input circuit.

DC amplifiers. Differential stage. Amplifying biopotentials requires amplifiers, whose bandwidth has lower limit $\omega = 0$. Amplifiers of this type are called ***DC amplifiers*** irrespective of what they amplify, be it current intensity, or voltage.

Analyzing the possibilities of using transistors in amplifier circuits (see § 22.4), one can think that amplification of slowly changing signals and DC signals does not differ from amplification of an alternating signal. As a matter of fact, explanation of the physical principles of transistor operation as amplifiers could be applied to DC as well. However, in case of negative feedback circuits, a circuit like one shown in fig. 22.16 would have a low amplifier gain, and it would be difficult to use one stage only. We have to use a series of cascades, but this leads to issues with amplifying slowly changing signals. The case is that in a DC amplifier, the stages should be connected without reactive elements (capacitors, transformers) that do not work with direct current. These circuit elements should be connected with conductors by galvanic way (fig. 22.20). Nevertheless, with such a connection, at the output of the stage, slow, random changes in voltage or current intensity (drift) will be amplified by the subsequent stages, which will lead to distortion of information.

Drift can be caused by wear of amplifier elements, effect of the temperature, instability of power supply voltage, etc. Thus, for direct amplification, it is

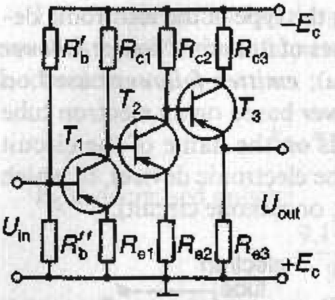


Fig. 22.20

necessary to try to reduce the drift in each stage and, in the input stage first of all.

One way to reduce the drift is using a differential cascade, the simplest diagram of which is shown in fig. 22.21; in this, a is T_1 and T_2; they are two identical transisto R_{c1} and R_{c2} are resistors in collector circui R_e is a resistor in emitter circuits; R is a variab resistor for tuning the circuit; R_l is the load resistor, from which output voltage U_{out} is picke up. Input voltage U_{in} is supplied on transist bases. The circuit shown in the figure is actually a bridge. In order to make thi clearer, let us present it in a conventional form for such circuits (fig. 22.21, b); let us split resistance R onto two shoulders, represent transistors T_1 and T_2 as a series connection of semiconductors with different mechanisms of electrical conductivity. A power supply is connected to one diagonal of the bridg and to the other diagonal resistor R_l of the load is connected.

The bridge is balanced, if the followi condition is valid:

$$\frac{R_{c1}+R/2}{R_{c2}+R/2}=\frac{R_{T1}}{R_{T2}}, \qquad (22.22)$$

where R_{T1} and R_{T2} are resistance of the first and second transistor respectively. In this ideal case, there is current through diagonal R_l and $U_{out}=0$. Possible changes in the power supply voltage or similar changes in the properties of transistors under the influence of temperatu and time will not disturb the bridge balanc and will not cause a current through resistor R_l. In a conventional DC amplifier, all such influences would inevitably lead to a drift.

The input signal will unbalance the bridge as the potentials of the bases of different transistors will become different, and R_{T1} and R_{T2} will change too.

Condition (22.22) will be violated, and in diagonal R_l of the unbalanced bridge, a curren will occur.

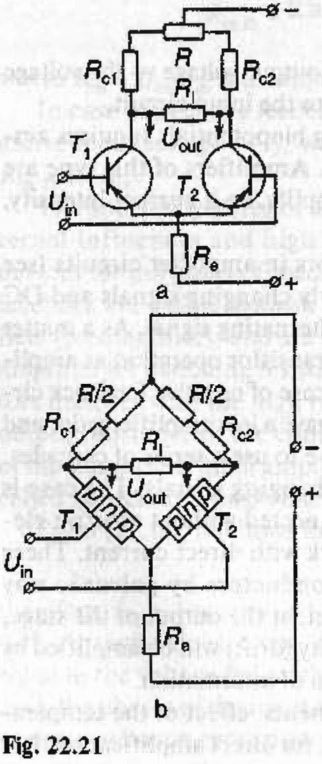


Fig. 22.21

The amplified output voltage will be proportional to the difference of input potentials on the bases, so such an amplifying stage is called a ***differential* stage**.

Since ideal balance is not attainable, there will inevitably be a slight drift in the differential stage. The balance of the circuit is improved adjusting variable resistor *R*.

A number of difficulties, which are specific to DC amplifiers, can be avoided by converting a slowly changing signal being amplified into an alternating voltage, whose amplitude is proportional to this signal.

Let us confine ourselves to considering a block diagram of such an amplifier with conversion (fig. 22.22). First, the amplified voltage (fig. 22.23, a) is converted into oscillations, whose amplitude varies over time in proportion to an input signal (fig. 22.23, b). After that, the resulting voltage is amplified (fig. 22.23, c) to the required value in a low frequency amplifier, which can be, for example, a resistance amplifier (see § 22.4). Further, this voltage is rectified to produce enveloping oscillations (electrical pulses) at the output. Thus, at the output of the amplifier, it is possible to generate voltage proportional to the input voltage, but much higher (fig. 22.23, d).

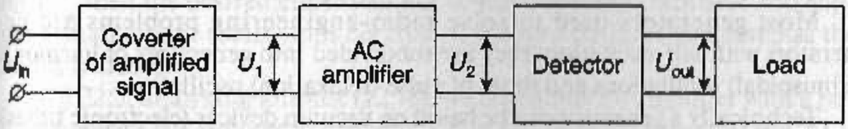


Fig. 22.22

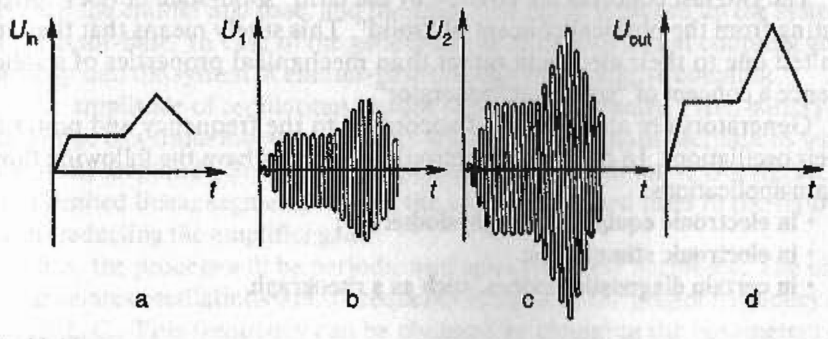


Fig. 22.23

Amplifiers of this type are notable for high stability, low drift and certain other advantages. No doubt, in addition to the foregoing circuits of DC amplifiers, there are a number of much more complicated devices with higher quality parameters.

Chapter 23

Generator

Generators (electronic generators) are devices that convert energy of DC voltage sources into the energy of electromagnetic oscillations of vary shape.

A large group of devices for use in medicine features structurally divers generators of electromagnetic oscillations. Along with generators, the chapter also describes the structure of an electronic oscillograph.

§ 23.1. VARIETIES OF GENERATORS OF ELECTRIC OSCILLATIONS

According to their operational principle, there are generators with self excitation (*self-oscillating systems* or *self-oscillators*) and those with extern excitation. They are essentially high-frequency power amplifiers.

Most generators used to solve radio-engineering problems are generators with self-excitation; they are subdivided into generators of harmo (sinusoidal) oscillations and those of pulse (relaxation) oscillations.

Technically a generator can be based on vacuum devices (electronic tubes gas-filled (discharge) tubes, semiconductor elements, and integrated circui

The two last concepts are covered by the term "solid-state devices" orig nating from the physical concept of "solid". This surely means that they are united due to their electricla rather than mechanical properties of a solid. Hence a concept of "solid-state generator".

Generators are also classified according to the frequency and power of their oscillations. In medicine, electronic generators have the following thre main applications:

- **in electronic equipment for physiotherapy;**
- **in electronic stimulators;**
- **in certain diagnostic devices, such as a rheograph.**

§ 23.2. TRANSISTOR-BASED GENERATOR OF HARMONIC OSCILLATIONS

Let us consider the operational principle of a transistor-based generat (fig. 23.1), producing self-oscillations close to sinusoidal ones. To understa

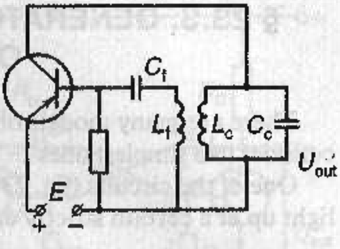


Fig. 23.1

the general operational principles of this generator, it is expedient to recall the block diagram of a self-oscillating system (see fig. 7.19).

Oscillating circuit L_0C_c is located in the collector circuit.

Inductor $L_{o.c.}$ coupled with L_c, serves as a feedback. The source of energy is battery ε. As a valve transmitting the energy into the circuit at the necessary time the transistor is used.

Upon switching on the circuit, in the oscillating circuit, small random electromagnetic oscillations occur. Due to inductive feedback, these oscillations are transmitted to the base, more exactly, to the p–n-junction between the emitter and the base, and amplified (see § 22.4). Oscillations, amplified by the transistor, are fed into the RLC through the collector circuit in resonance with those that already exist there. The amplitude of oscillations increases. Surely, this will happen only at a certain phase interrelation between the oscillations in the circuit and the change in the voltage on the base. The feedback should be positive. If we change the ends of coil winding $L_{o.c}$ then the desired effect will not be achieved: the transistor will suppress small circuit oscillations because of random currents generated as the circuit is closed.

In § 22.5, analyzing formula (22.18), we noted that an amplifier with a positive feedback can become a generator. Let us compare the generator (fig. 23.1) with an amplifier (fig. 22.8, b). In both cases, a signal to be amplified is fed between the emitter and base, and the amplified signal is picked up the system of collector-base. In case of the generator the amplified signal coupling gets partially into the system of emitter-base through the inductive coupling.

The amplitude of oscillations cannot increase incessantly. First, battery E is a source of a finite amount of energy and cannot maintain oscillations with an infinite amplitude. Second, the amplitude characteristic curve (see fig. 22.2) has a limited linear segment; leaving the segment (dashed lines in the figure) means reducting the amplifier gain.

Thus, the process will be periodic and approximately harmonic. The circuit generates oscillations whose frequency is equal to the proper frequency of circuit L_cC_c. This frequency can be changed by changing the parameters of the circuit, which are its inductance and capacitance. For structural reasons, capacitor C_c is usually made to be variable. Circuit elements R_b and C_b serve to create the desirable voltage biasing at the base so as to select the optimal position of the operating point (see fig. 22.2, point A on the curve) and not to induce a short-circuit of voltage source ε with the emitter-base circuit.

§ 23.3. GENERATORS OF PULSE (RELAXATION) OSCILLATIONS

There are many models of generators of relaxation oscillations, but let u consider two simplest ones.

One of the circuits (fig. 23.2) is a generator using neon tube *T*. Such tub light up at a certain strictly defined voltage U_1 and switch off at a lower vol age U_g. The process starts with chargi capacitor *C*.

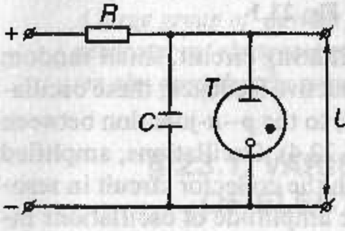


Fig. 23.2

On the curve of the dependence of th output voltage on time (fig. 23.3), thi stage is shown with segment *OA* corr sponding to equation (18.20). At point *A*, the voltage at the capacitor attains U_g which is sufficient to ionize the gas insi the neon tube; it lights up, and the cap citor discharges through it [see (18.19)]. At point *B*, the voltage at the tub will be equal to U_g, the tube goes off, and its resistance increases very muc The capacitor recharges and the process is reiterated.

As we can see from (18.20), in such a circuit, the rate of voltage increa can be changed, changing parameters *R* and *C*. For instance, an increase in resistance will lead to an increase in time *T*, segment *OA* will become mor gently sloping. At segment *AB*, the voltage changes as the neon tube is di charging and, therefore, depends on its characteristic curves. Adjusting pa rameters of the circuit, we can make the real curve (fig. 23.3) closer to th ideal one; it is called ***saw-tooth voltage*** (fig. 23.4). The curve of this voltage de pendence on time is like the profile of a saw. During time *T*, the voltage grow from U_1 to U_2 linearly; then, during time τ, it decreases to its minimum valu also linearly. If a more accurate approximation of the linear increase of th voltage over time is required, more sophisticated circuits are used. Saw-too

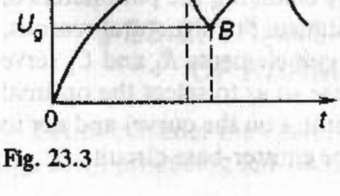


Fig. 23.3

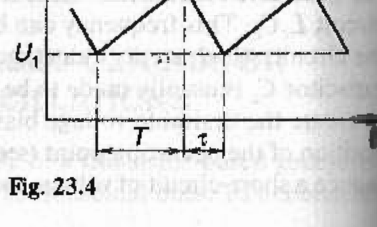


Fig. 23.4

voltage is used in the sweep of an electronic oscilloscope (see § 23.4).

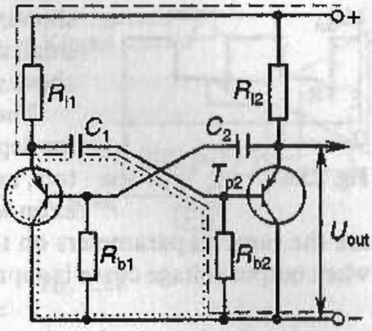


Fig. 23.5

The generator of relaxation oscillations is also a multi-vibrator, one of its circuits is shown in fig. 23.5. To understand the operation of this circuit, let us recall briefly the property of a transistor to pass or not to pass current depending on the voltage on the base. Resistor R_{b1} is connected in parallel to *p*–*n*-junction collector-base of the first transistor. If a current flows through this resistor from top to bottom, then the potential of the transistor base is higher than the potential of the collector. This circumstance increases the electrical conductivity of the transistor and the current can flow through it from the emitter to the collector. When the current passes through resistor R_{b1} in the opposite direction, i.e. from bottom to top, the base potential will be less than the collector potential, and the transistor will be cut off.

To begin our explanation, let us assume that transistor T_{p1} is cut off, the source charges capacitor C_1 (the circuit is indicated by shading). After capacitor C_1 has charged up to a certain voltage, it will, with active transistor T_{p1}, discharge through it (the circuit is shown by dotted line). Similar reasoning can be offered for capacitor C_2. It is important to note that each capacitor is discharged through the next resistor, thus affecting its state.

So far this process is quite similar to operation of the relaxation generator with a neon tube where the capacitor was charged through resistor *R*, here it is charged through R_{n1} and R_{b2}, and discharged through neon tube *T*; here the capacitor does it through transistor T_{p1} and resistor R_{b2}. However, in the previous case, the current passed through the tube at a certain voltage value; here it depends on the voltage drop at R_{b1} due to charging and discharging of capacitor C_2. While capacitor C_2 is charged (through R_{n2} and R_{b1}), the current through R_{b1} flows from top to bottom, and transistor T_{p1} is active. As this capacitor is discharging (through T_{p2} and R_{b1}), the current through R_{b1} is directed from bottom to top, and, at its correspondent value, transistor T_{p1} is cut off. Thus, as one capacitor is charged, the other is discharged; when one transistor is active, the other is cut off, and vice versa. The process is repetitive.

Surely, we imagine a case when both transistors are active, and the voltages and current intensities on all elements are constant. A more detailed study

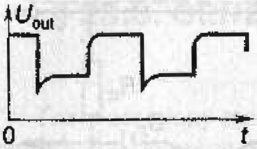


Fig. 23.6

shows that such a state will be unstable, and any small deviation from it will lead to jumps describ above.

The nature of output voltage U_{out} of multi-vi tor depends on the parameters of resistors, cap tors, and transistors included in the device, as we as on whether the parameters of elements on the left are the same as parameters on the right, etc. Let us discuss a possible cas when output voltage curve is approximately rectangular (fig. 23.6).

§ 23.4. ELECTRONIC OSCILLOSCOPE

An ***oscilloscope*** is a measuring instrument for visual observation or recor ing of a functional dependence between two quantities converted into an elec trical signal.

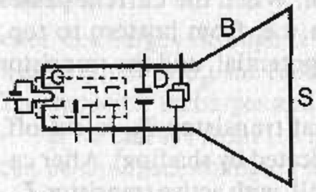


Fig. 23.7

Oscilloscopes are widely used to observe th time dependence of a variable.

The main part of the electronic oscillos is its cathode-ray tube (fig. 23.7). Its elemen are located in vacuum bulb *B*. They inclu fluorescent screen *S*, deflection system *D* con sisting of capacitors, and electron-beam gun (shown by dashed line) consisting of a heat cathode, which is similar to the cathode of a diode, and special electrodes that accelerate and focus electrons. To the capacitor plates, a difference of pot tials is applied. Depending on its sign and value, the electron beam is deflect ed in the vertical or horizontal direction. The electron beam, formed and di rected in a certain way, falls onto the luminescent screen — the front wa of the cathode ray tube covered with luminophores, which are able to glow be cause under the impact of electrons (***cathodoluminescence***).

The electron beam will be represented on the screen as a glowing point. Changing the voltage on the deflection plates gradually, the glowing point can be moved about the screen. Luminophores have a property of afterglow: they glow for some time in a certain place after the electron beam moved away from this place. Therefore, the beam movement is visible on the screen in the form of a line.

A block diagram of the oscilloscope is given in fig. 23.8: Amp. *X*, Am *Y* are amplifiers; PS is a power supply unit; SG is a sweep generator; EB is an electron beam tube. There is also a synchronization unit. Fig. 23.9 show the front panel of the oscilloscope.

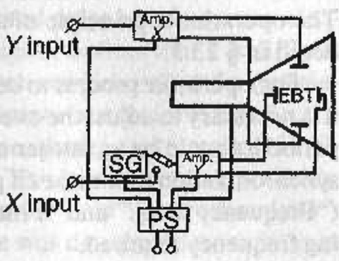

Amp. X
Y input
EBT
SG
Amp. Y
X input
PS

Fig. 23.8

The signal supplied to terminals *Y*-input and GRND is amplified and fed onto vertical deflection plates. On the oscilloscope screen, a segment of a vertical line will represent such a signal.

To watch the time dependence of the signal, the luminous point should move simultaneously and uniformly in the horizontal direction. In order to record a periodic process, the point should move to the right of the screen over a finite time interval and return in the shortest possible time. That is why, voltage applied to the horizontally deflecting plates, should have a saw-tooth shape (see, for instance, fig. 23.4, and $T \gg \tau$).

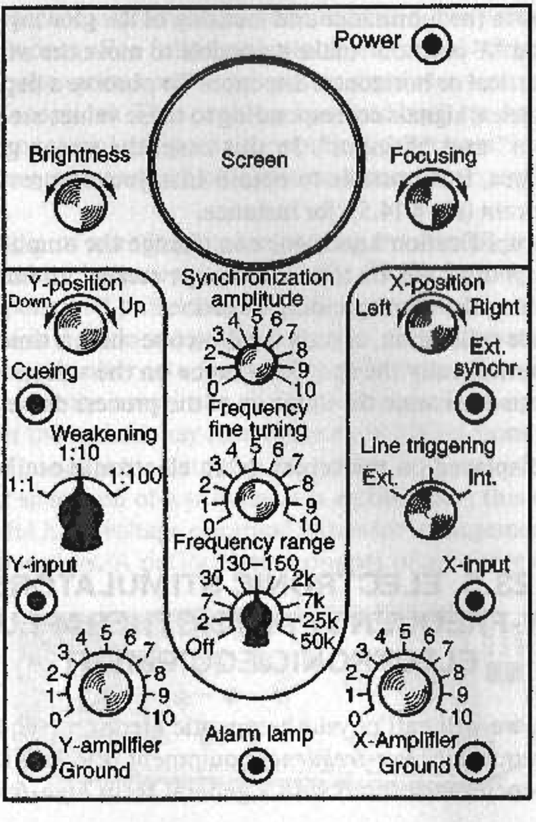

Power
Brightness
Screen
Focusing
Y-position
Down
Up
Synchronization amplitude
0 1 2 3 4 5 6 7 8 9 10
X-position
Left
Right
Ext. synchr.
Cueing
Frequency fine tuning
0 1 2 3 4 5 6 7 8 9 10
Weakening
1:1
1:10
1:100
Line triggering
Ext.
Int.
Frequency range
Off
2
7
30
130–150
2k
7k
25k
50k
Y-input
X-input
0 1 2 3 4 5 6 7 8 9 10
Y-amplifier
Ground
Alarm lamp
0 1 2 3 4 5 6 7 8 9 10
X-Amplifier
Ground

Fig. 23.9

The operational principle of a device serving as sweep generator was considered in § 23.3.

For a periodic process to be displayed on the screen as an immobile image, it is necessary to adjust the sweeping frequency quite accurately: per one sweep period it should be an integer number of periods of the studied signal. A sweep synchronization unit makes it possible to observe this condition. Twist handles ("Frequency range" and "Fine tuning") make it possible to set up the sweeping frequency required.

If the process being studied is single or non-periodic, the standby sweep mode, provided in certain oscilloscopes, can be used. This sweep mode is active every time only when a recorded process takes place.

Rotating "Brightness" and "Focus" twist handles, the difference of potentials between the accelerating electrodes is changed, thanks to which, it can set up different intensity and cross-sectional area of the electron beam. This leads to changes in the luminance and focusing of the glowing point. Handles "*Y*-position" and "*X*-position" make it possible to move the whole picture displayed in the vertical or horizontal direction. To observe a dependence of any two values, electrical signals corresponding to these values are supplied to terminals "*Y*-input" and "*X*-input". In this case, the sweep generator is not switched on. Thus, it is possible to obtain Lissajous figures (see § 7.3) and a vector cardiogram (see § 14.5), for instance.

Using the amplification knob, one can change the amplification gain of studied signal. In this case, on the oscilloscope screen, an image is stretched or contracted along the corresponding direction.

For time scale calibration, certain oscilloscopes have a time marks generator to change periodically the spot luminance on the screen. Thanks to this property, one can determine the duration of the process displayed or its separate stages.

An image displayed on the screen of an electronic oscilloscope can be photographed.

§ 23.5. ELECTRONIC STIMULATORS. LOW-FREQUENCY PHYSIOTHERAPEUTIC ELECTRONIC EQUIPMENT

For brevity, we will call physiotherapeutic electronic equipment of low and acoustic frequencies ***low-frequency*** equipment. Electronic equipment of all other frequencies is covered with a general term ***high-frequency*** equipment.

Medical devices — generators of harmonic and pulse low-frequency electromagnetic oscillations — unite two large groups of devices that it is difficult to distinguish clearly: ***electronic stimulators*** (***electrostimulators***) and ***physiotherapeutic devices***. At low frequencies, the most important factor is the electrical, rather than action of the current. Therefore, therapy with current is stimulation of effect using irritation by current. This circumstance, probably, also caused a mix-up of concepts "therapeutic device" and "electrostimulator".

Although electrical stimulation of a muscle was discovered as early as in the 18th century, widespread use of electric stimulators began only in recent decades. Nowadays, there are quite a lot of different electrostimulators. However, even at present, the researchers face important medical and physiological problem of accurate setting of the output parameters of an electrical signal of electric stimulators: pulse shape, its duration, pulse current frequency, and its duty cycle (see Chapters 18 and 19).

Nowadays, electric stimulators can be subdivided into ***permanent***, ***portable***, and ***implantable*** ones. For fully implantable electrical stimulators, such as pacemakers, a quite serious problem is power supply sources, which should be durable and economical. This problem is solved both by developing appropriate power sources and designing economical generators. For example, it is desirable to have generators, which consume almost no energy in the interval between pulses.

As an example of a multipurpose permanent stimulator, let us look at a universal electric pulse generator UEI-1 (fig. 23.10). It is a generator of rectangular and exponential pulse current. Pulse parameters and their frequency can be adjusted in a wide range, for example, the duration of rectangular pulses can discretely alternate from 0.01 to 300 msec. The device makes it possible to measure the amplitude of a current pulse in the patient's circuit. On the screen of the cathode ray tube (left side of UEI-1 front panel), you can see the shape of pulses at the output of the device.

A particular specimen of a stimulator is ***defibrillator***; this device is a generator of powerful high-voltage electrical pulses for management of severe cardiac rhythm disorders. A defibrillator consists of an energy accumulation

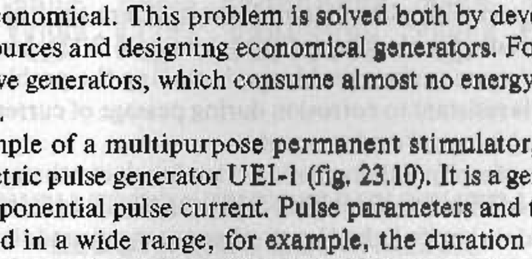


Fig. 23.10

device (capacitor), capacitor-charging device, and discharge circuit. Fig. 23.11 shows the outward appearance of pulse defibrillator DI-03.

An implantable radio frequency pacemaker ECSR-01 (fig. 23.12) is a porta and often implantable device. Its implantable part (receiver) is shown in the middle of the figure, it weighs 22 g, it is 8.5 mm thick. The receiver receives radio nals from an external transmitter (on the left of the figure). These signals are ceived inside the patient's body by the implantable part and are transmitted in the form of pulses to the heart through electrodes. In the right part of the figur we see power supply unit, which like the transmitter, is worn by the patient on the outside.

A special kind of electrostimulators are those that are able to transmit information in a coded form, usually perceived by sense organs. Such a stimulator is a cochlear prosthesis, which converts acoustic information into an electrical signal, i.e. essentially replaces the cochlea of the inner ear (see § 8.5). A portable cochlear prosthesis is shown in fig. 8.13.

Technical devices of electric stimulation also include electrodes for supplying an electrical signal to a biological system. In many cases, electric stimulation is achieved with plate electrodes, which are applied on the huma body like electrodes for electrocardiography (see § 21.2). Implantable el trodes generate more serious problems, including the problem of choo a material that is resistant to corrosion during passage of current in conditio of an aggressive biological environment.

An example of a physiotherapeutic device for electrotherapy with sinusoidal modulated currents is Amplipuls-3 (fig. 23.13, a). In it, the frequency of carrier sinusoidal signals is 5 kHz, the frequency of modulating sinusoid oscillations can be gradually regulated within the range of 10–150 Hz. Som possible forms of currents, generated by this device, are shown in fig. 23.13, b; the proportion between the frequencies of the carrier and modulating oscillations is not observed in the figure.

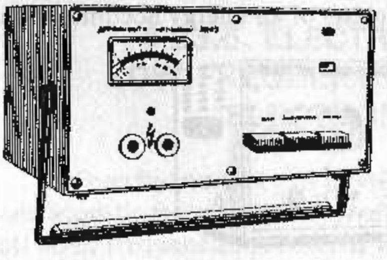


Fig. 23.11

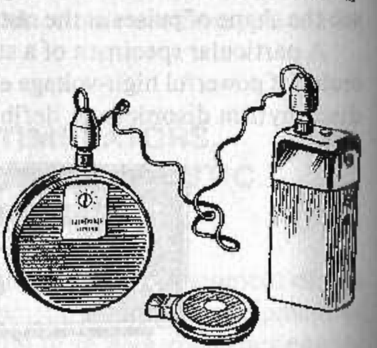


Fig. 23.12

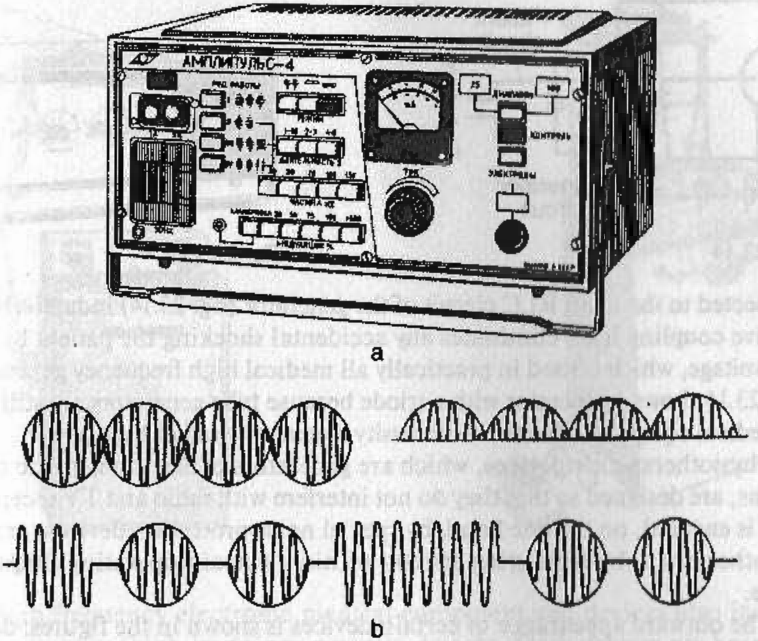


Fig. 23.13

§ 23.6. HIGH FREQUENCY PHYSIOTHERAPEUTIC ELECTRONIC EQUIPMENT. ELECTROSURGERY DEVICES

A large group of medical devices — generators of electromagnetic oscillations and waves — operates in the range of ultrasonic (ultra-audible), high, ultrahigh and microwave frequencies and is covered by a generalizing term ***high frequency electronic equipment.***

The problem of electrodes is solved in this case in different ways. For high frequency currents, glass electrodes are used (see fig. 19.5); the impact of alternating magnetic field (inductothermy) is implemented through spirals or flat folded cables carrying alternating current. For UHF-therapy, the body part to be heated is placed between disc-shaped metal electrodes (fig. 23.14) covered with an insulating layer. For exposure to electromagnetic waves, their emitter is placed closer to the body.

For the patient's safety, the electrodes are connected not to the RLC circuit of the generator, but to the patient's circuit (therapeutic circuit), which is

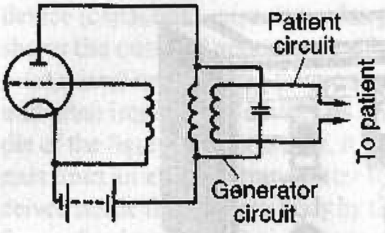


Fig. 23.14

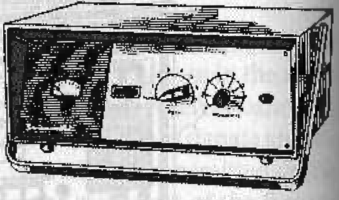


Fig. 23.15

connected to the main RLC circuit of the generator (fig. 23.14) inductively. Inductive coupling itself eliminates any accidental shocking the patient by hig DC voltage, which is used in practically all medical high frequency generato Fig. 23.14 shows a generator with a triode because tube generators are still used in medical equipment due to the necessity to generate rather high power.

Physiotherapeutic devices, which are generators of electromagnetic oscillations, are designed so that they do not interfere with radio and TV receptio This is ensured, on the one hand, by special noise protection devices, and, on the other hand, by remaining strictly within a certain operating frequen range.

The outward appearance of certain devices is shown in the figures: devic Iskra-1 is a high-frequency generator operating in pulse mode and used for local darsonvalization (fig. 23.15). Device IKV-4 is intended for inductother and operates at a frequency of 13.56 MHz (fig. 23.16). For UHF-therapy, the is a portable device UVCh-66 (fig. 23.17). There is also a device for microwa therapy "Luch-58" (fig. 23.18).

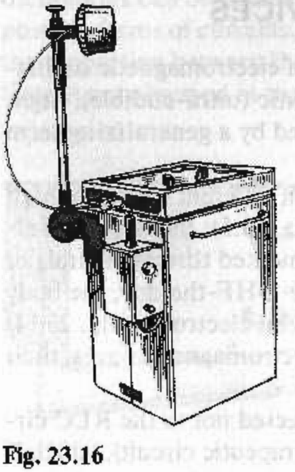


Fig. 23.16

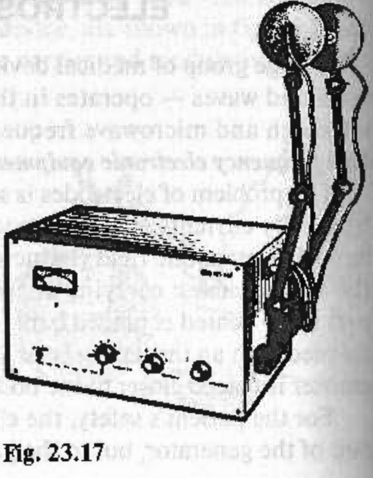


Fig. 23.17

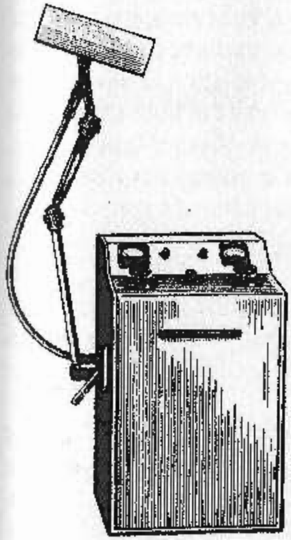


Fig. 23.18

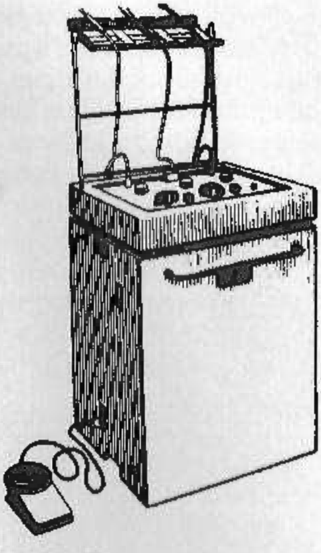


Fig. 23.19

High frequency electronic medical equipment and devices also include electrosurgery (high frequency surgery) devices. Such devices are based on a generator of harmonic or modulated electromagnetic oscillations. The power of electromagnetic oscillations used in electrosurgery can be from 1 W to several hundred Watts.

A specific feature of generators is the fact that they should put out their power on the load (biological tissue) which varies within wide limits. For a long time, generators can operate without any load at all, that is why, in the electrosurgery devices, vacuum tubes are still widely used. The tubes, in comparison with semiconductor devices, are more resistant to possible overloads.

In electrosurgery, electromagnetic oscillations are supplied on electrodes that cut or coagulate a tissue. There are electrodes for monopolar and bipolar electrosurgery.

In the first case, one output of the device generator is connected to the active electrode, which is used for electrosurgery, and the other electrode (passive one) contacts with the patient's body.

In the second case, both outputs of the generator are connected to two active electrodes, between which a high-frequency current flows and exerts a surgical action. In this case, both electrodes are active, and the passive one is not used.

The outward appearance of electrosurgery devices is shown in fig. 23.19.

Section 6

OPTICS

Optics is for branch of physics dealing with laws of light emission, absor tion, and propagation. In physics, the term "light" is used not only for radi tion perceived by the human eye, but also for invisible radiation. The nature of light is two-old (dual). It means that light exibits the traits of both an electr magnetic wave and a flow of particles — photons. The dualism of light, in pa ticular, is expressed in formula $\varepsilon = h\nu$ because photon energy e is a quantu characteristic, and oscillation frequency ν is a characteristic of a wave proces In some optical phenomena, light acted as a wave, and in others, as a particl. The dual nature is inherent to particles as well: electrons, protons, etc. Sinc light has electromagnetic properties, it is expedient to study optics after elec trodynamics. Problems of light emission are on borderline with atomic physi and are essentially connected with it. That is why, the section Optics preced one on atomic physics.

In the development of physics, optical experiments and theories were of a special importance: laws of rectilinear light propagation and its reflecti from mirror surfaces were one of the first physical laws known long befor Christ; Michelson's interference experiment provided experimental eviden of the theory of relativity; Planck's hypothesis about special-energy distrib tion of radiation was the beginning of quantum physics. Studies of visible ligh and its measurement are not only the province of physics, but also of physi logy. In this respect, optics is like acoustics. For physicians and biologists, it is first important for studying biological objects: microscopy, spectromet refractometry, polarimetry, calorimetry. In addition, physicians should kno the physical principles of thermal radiation in diagnostics of diseases (therm graphy), the structure of equipment for light therapy, and other importa issues.

Chapter 24

Interference and diffraction of light. Holography

Interference of light is a combination of light waves forming a stable picture of their amplification and attenuation. To induce light interference, certain conditions should be met.

Light diffraction is the phenomenon of light deflection from rectilinear propagation in a medium with marked inhomogeneities. The possibility to observe difrraction depends on the ratio between the light wavelength and the size of inhomogeneities. Diffraction of spherical waves (Fresnel diffraction) and plane-parallel waves (Fraunhofer diffraction) are distinguished with a certain degree of conventionality. Description of a diffraction pattern is possible, taking into account interference of secondary waves.

The chapter deals with holography as a method based on interference and diffraction.

§ 24.1. COHERENT LIGHT SOURCES. CONDITIONS FOR MAXIMUM AMPLIFICATION AND ATTENUATION OF WAVES

The interference of waves propagating in the medium is determined by a combination of corresponding oscillations. The simplest case of electromagnetic wave interference is observed when their frequencies are equal and their electric vectors are parallel. In this case, the amplitude of the resulting wave can be found by formula (7.20), which for the electric field intensity is written as:

$$E^2 = E_1^2 + E_2^2 + 2E_1E_2\cos\Delta\varphi, \qquad (24.1)$$

where $\Delta\varphi$ is the phase difference of waves being combined (oscillations).

Depending on the type of light sources, the result of wave interference can be fundamentally different.

First, let us consider interference of waves coming from ordinary light sources (bulb, flame, the sun, etc.). Each such source is a totality of a large number of radiating atoms. A separate atom emits an electromagnetic wave in approximately 10^{-8} s, and emission is a random event, so, in formula (24.1), phase difference $\Delta\varphi$ takes random values correspondently. For all that, the average value of $\cos\Delta\varphi$ in emission of all atoms is equal to zero. Instead of (24.1), we obtain an averaged equation for those points of space where two waves going from two usual light sources are interfering:

$$\langle E^2\rangle = \langle E_1^2\rangle + \langle E_2^2\rangle. \tag{24.2}$$

Since the intensity of wave is proportional to the square of its amplitude, then from (24.2), we have the condition of adding wave intensities I_1 and I_2:

$$I = I_1 + I_2. \tag{24.3}$$

It means that for the intensities of emissions going from two (or more) usual light sources, a quite simple addition rule operates: the intensity of the total emission is equal to the sum of the intensities of interfering waves. This can be seen in everyday practice: illumination from two bulbs is equal to the sum of illuminations generated by each bulb separately.

If $\Delta\varphi$ is constant, there is interference of light. In different points of space, the intensity of the resulting wave has values from the minimum to maximum ones.

Interference of light occurs if the light sources are coordinated (*coherent*) which provides a constant phase difference $\Delta\varphi$ of interfering waves at different points. Waves that meet this condition are called *coherent* ones.

Interference could be induced from two sinusoidal waves of equal frequency, but it is practically impossible to generate such light waves, so coherent waves are obtained by splitting a light wave coming from the source.

This approach is used in the *Young's experiment*. An opaque obstacle with two slits is placed in the path of a spherical wave coming from source *S* (fig. 24.1). The points of the wave surface that reached the obstacle become the centers of coherent secondary waves, so the slits can be considered as coherent sources. Interference is observed on screen *Sc*.

Another method consists in creating a virtual image *S'* of source *S* (fig. 24.2) with a special single-layer mirror (Lloyd's mirror). Sources *S* and *S'* are coherent. They provide conditions for wave interference. The figure shows two interfering beams hitting point *A* on screen *Sc*.

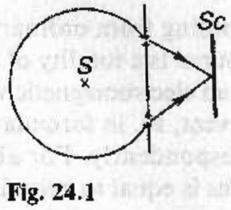


Fig. 24.1

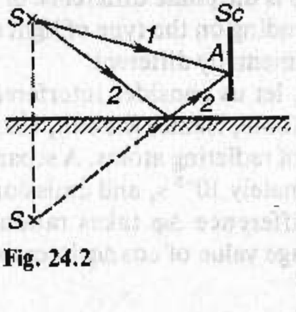


Fig. 24.2

While emission time τ of a separate atom is finite, then, for interference, path difference δ of beams *1* and *2* cannot be too great; otherwise, different non-coherent waves would meet at point *A*. For interference, the maximum value of δ is calculated by the speed of light and the time of atom emission:

$$\delta = c\tau = 3 \cdot 10^8 \cdot 10^{-8} = 3 \text{ m}. \tag{24.4}$$

The interference pattern can be quantified using formula (24.1) if we already know the phase difference of the interfering waves and their amplitude.

The following particular cases are of practical interest: maximum wave amplification — maximum intensity (*max*); the highest attenuation — minimum intensity (*min*).

Note that it is more convenient to express conditions of maxima and minima of intensities through the path difference rather than phase difference; as for interference, the paths passed by coherent waves are usually known. Let us show it with an example of interference of plane waves I, II whose vectors $\overline{E}$ are perpendicular to the drawing plane (fig. 24.3).

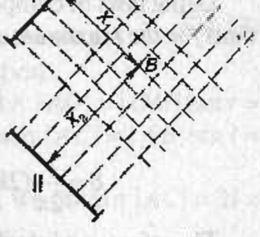


Fig. 24.3

Oscillations of vector $\overline{E}$ of these waves at a certain pont *B* placed at distances x_1 and x_2 from each source respectively, meet the following harmonic law:

$$E_1 = E_{\max 1} \cos \omega(t - x_1/\upsilon_1), \quad E_2 = E_{\max 2} \cos \omega(t - x_2/\upsilon_2). \tag{24.5}$$

To generalize our derivation, let us assume that the waves propagate in different media[1] with refraction indices n_1 and n_2. The velocities of wave propagation are equal to $\upsilon_1 = c/n_1$ and $\upsilon_2 = c/n_2$, respectively, where c is the speed of light in vacuum. Then from (24.5), we can obtain an expression for the phase difference[2]:

$$\begin{aligned} \Delta\varphi = \varphi_2 - \varphi_1 = \omega(t - x_2/\upsilon_2) - \omega(t - x_1/\upsilon_1) = \\ = \omega(x_1/\upsilon_1 - x_2/\upsilon_2) = (2\pi/(Tc))\,(x_1 n_1 - x_2 n_2). \end{aligned} \tag{24.6}$$

Since in vacuum, wavelength $\lambda = Tc$, instead of (24.6), we have

$$\Delta\varphi = (2\pi/\lambda)(x_1 n_1 - x_2 n_2). \tag{24.7}$$

---

[1] The schematic nature of fig. 24.3 does not permit showing different media for propagation of different waves.

[2] It is useful to note that since $\Delta\varphi$ does not depend on time, the interfering waves are coherent.

The product of the geometric path of the wave by the refraction index of the medium, that is xn, is called ***optical path length***[1], and the difference of these paths

$$\delta = x_1 n_1 - x_2 n_2 \qquad (24.$$

is called ***optical path difference of interfering waves.***

Based on (24.7) and (24.8), we obtain a relationship between the phase difference and the optical path difference of interfering waves:

$$\Delta\varphi = \frac{\lambda}{2\pi}\delta \quad \text{or} \quad \delta = \frac{\lambda}{2\pi}\Delta\varphi. \qquad (24.$$

Using laws of composition of oscillations and ratio (24.9), we obtain ***conditions for the maximum*** and ***minimum*** light intensity for interference:

$$\delta = \frac{\lambda}{2\pi}2k\pi = k\lambda = 2k\cdot\frac{\lambda}{2}, \qquad (24,10)$$

$$\delta = \frac{\lambda}{2\pi}(2k+1)\pi = (2k+1)\frac{\lambda}{2}, \quad (k = 0,1,2\ldots). \qquad (24.11)$$

Therefore, **an interference maximum is present at points where the optical path difference is equal to an integer number of wavelengths** (even number of half waves); **the minimum is present at points where the optical path difference is equal to an odd number of half-waves.**

§ 24.2. THIN PLATE (FILM) INTERFERENCE. ANTI-REFLECTION COATING

Formation of coherent waves and interference also take place when light hits a thin transparent plate or film. A streak of light falls onto a plane-parallel plate (fig. 24.4). From this streak of light, beam *1* falls onto point *A* and is reflected partially (beam *2*) and refracted partially too (beam *AM*). The beam refracted ray is reflected on the lower surface of the plate at point *M*. After refraction at point *B*, the reflected beam passes into the first medium (ray *3*). Beams *2* and *3* are formed from one beam, that is why they are

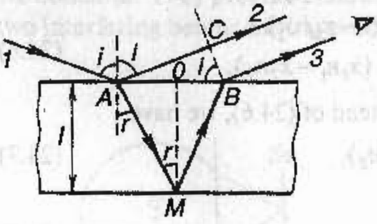


Fig. 24.4

---

[1] Segments with the equal optical path length, i.e. segments that are covered by light over the same time are called *tautochronic* ones.

coherent and will interfere. Let us find the optical path difference for beams *2* and *3*. To do this, from point *B*, we draw normal *BC* to the beams. From straight-line *BC* to the meeting point of the beams, their optical path difference will not change, a lens or eye will not add any phase difference.

Before divergence at point *A*, these rays together with others not shown in Fig. 24.4 formed beam *1* and, surely, had an equal phase. Beam *3* passed the distance $|AM| + |MB|$ inside the plate with refraction index n, beam *2* passed distance $|AC|$ in the air. Therefore, their optical path difference is:

$$.\delta = (|AM|+|MB|)n - |AC| = 2|AM|n - AC|, \tag{24.12}$$

since $|AM| = |MB|$. According to the reflection law,

$$n = \sin i/\sin r \quad \text{or} \quad \sin i = n\sin r, \tag{24.13}$$

where i is the angle of incidence; r is the angle of refraction.

From ΔAMO, we have: $|AM| = |OM|/\cos r = l/\cos r$, $|AO| = |OM|\tan r = l\tan r$ (l is plate thickness). From ΔACB, we obtain $|AC| = |AB|\sin i = 2|AO|\sin i$.

Taking into account these equations and (24.13), we obtain $|AC| = 2l \times \tan r \cdot n \cdot \sin r = 2ln \cdot \sin 2r/\cos r$.

Then,

$$\delta = 2ln/\cos r - 2ln/\sin^2 r/\cos r = 2ln\cos r.$$

Thus, the optical path difference of the beams is

$$\delta = 2ln\cos r. \tag{24.14}$$

Formula (24.14) does not take into account one important circumstance. Practice shows that during light reflection from a more optically dense medium, that is one with a higher refraction index, the wave phase changes by π, which corresponds [see (24.9)] to a change in optical path difference by $\lambda/2$, that is **in case of reflection from a more optically dense medium, a *half-wave is lost***[1].

If both beam *2* and *3* lost a half-wave, it would not change expression (24.14). However, beam *2* is reflected from a more optically dense medium (point *A*) and loses a half-wave, but beam *3* is reflected from a less optically dense medium (point *M*), and its phase is not changed. Taking into account the half-wave loss, the optical path difference is:

$$\delta = 2ln\cos r - \lambda/2, \quad \text{or} \quad \delta = 2ln\cos r + \lambda/2. \tag{24.15}$$

---

[1] For cyclic processes, it does not matter whether the phase decreases or increases by π, therefore, so one could also speak not about loss, but about an increment of a half-wave; however, such terminology is not in current use.

Since $\cos r = \sqrt{1 - \sin^2 r} = \sqrt{1 - (\sin^2 i)/n^2} = \sqrt{(n^2 - \sin^2 i)/n^2}$, then δ can be also expressed through the angle of incidence:

$$\delta = 2ln\sqrt{(n^2 - \sin^2 i)/n^2} - \lambda/2 = 2l\sqrt{n^2 - \sin^2 i} - \lambda/2. \qquad (24.16)$$

For the ***maximum of interference*** [see (24.10), (24.16)], we obtain:

$$2l\sqrt{n^2 - \sin^2 i} = (2k+1)(\lambda/2). \qquad (24.17)$$

For the ***minimum of interference*** [see (24.11), (24.16)], we obtain:

$$2l\sqrt{n^2 - \sin^2 i} = k\lambda\text{[1]}. \qquad (24.18)$$

Formulas (24.17) and (24.18) correspond to interference in reflected light too. Interference of light passing through the plate is shown in fig. 24.5 showing only those beams, which are necessary for understanding of the phenomenon.

The reader can independently derive the corresponding formulas and make sure that for this case (24.17) corresponds to the minimum of interference, and (24.18) does to its maximum. In the light of the energy conservation law this is understandable, since interference is redistribution of light energy: the plate redistributes the inclination flux into reflected and passing fluxes (in this, we neglect the absorption). At all that, if the reflected flux is maximum, the passing one is minimum, and vice versa.

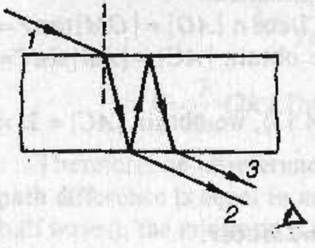


Fig. 24.5

In case of reflection case, interference is present more clearly than in the passing light due to the sizeable difference in the intensities of the reflected and passing beams. If we assume that about 5% of the incident energy is reflected at the interface of two transparent media, then

$$I_2 = 0.05 I_1, \qquad (24.19)$$

where I_1 and I_2 are intensities of beam *1* and *2* respectively (see fig. 24.4). The intensity of beam *3* (taking into account double refraction and single reflection) is equal to

$$I_3 = 0.95 \cdot 0.05 \cdot 0.95 I_1. \qquad (24.20)$$

From (24.19) and (24.20), we obtain:

$$I_2 : I_3 \approx 1.1, \qquad (24.21)$$

---

[1] At maxima and minima, in order for k to have the same values (0, 1, 2, etc.) into the formula for δ, we write $(c + \lambda/2)$ instead of δ.

which means approximate equality of amplitudes of interfering beams upon reflection: the condition of minimum corresponds to almost absolute darkness. Making similar calculations for passing light (fig. 24.5), we obtain:

$$I_2 = 0.95 \cdot 0.95 I_1, \quad I_3 = 0.95 \cdot 0.05 \cdot 0.05 \cdot 0.95 I_2;$$
$$I_2 : I_3 \approx 400,$$

or for amplitudes,

$$A_2 : A_3 \approx 20. \tag{24.22}$$

From (24.22), it follows that in passing light, waves with essentially different amplitudes interfere, so, maxima and minima differ from each other little, and interference is hardly noticeable.

Let us analyze dependencies (24.17) and (24.18). If a parallel beam of monochromatic radiation falls onto a plane-parallel plate at a certain angle, then, according to these formulas, the plate looks bright or dark in the reflected light.

When the plate is illuminated with white light, the maximum and minimum conditions are met for individual wavelengths; the plate will become colored; for all that, the colors in the reflected and passing light will complement each other creating white color.

In real-life conditions, the incident beam is not strictly parallel and does not have one definite angle of incidence *i*. At sizeable thickness *l* of plate, such small scattering of *i* can lead to sizeable difference of the left member in formulas (24.17) and (24.18), and maximum and minimum conditions will not be observed for all beams of the light streak. This is one of the reasons why interference can only be present in thin plates and films.

When monochromatic light falls onto a plate of variable thickness, to each value of *l*, its own interference condition corresponds, so the plate is crossed by light and dark lines (strips): ***lines of equal thickness***. Thus, in a wedge, this is a system of parallel lines (fig. 24.6); in an air gap between a lens and plate, such lines are rings (*Newton's rings*).

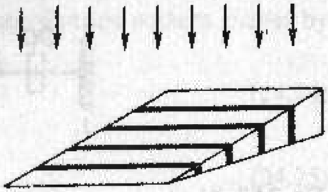


Fig. 24.6

When white light falls onto a plate of variable thickness, multi-colored spots and lines occurs. We can see colored soap films, films of oil and lubricant on the water surface, irridescent colors of certain insects and bird wings. In these cases, of the films are not always completely transparent.

Thin-film interference is of particular practical interest in connection with designing devices that reduce the proportion of light reflected by optical systems and therefore, increasing energy received by recording systems — photographic plates, eye, etc. For this purpose, the surface of an optical system is covered with a thin layer of metal oxide so that for a certain wavelength average for this spectral region, there is minimum interference in reflected light. As a result, the portion of passing light increases. Coating opt cal surfaces with special films is called antireflection coating, and optic articles themselves with such coatings are called *coated optics*.

If a specially chosen number layers are applied onto glass surface, it is possible to create a reflective light filter, which, due to interference, will pass or reflect a definite wavelength range.

§ 24.3. INTERFEROMETERS AND THEIR APPLICATION. NOTION OF INTERFERENCE MICROSCOPE

Light interference is used in special devices — *interferometers* — to measure wavelengths with a high degree of accuracy, short distances, refracti indices of materials, and to control the quality of optical surfaces.

Fig. 24.7 shows a schematic block diagram of a *Michelson interferome* which is a twin-wave interferometer as in it, a light wave bifurcates[1] and its both parts interfere after passing the different paths.

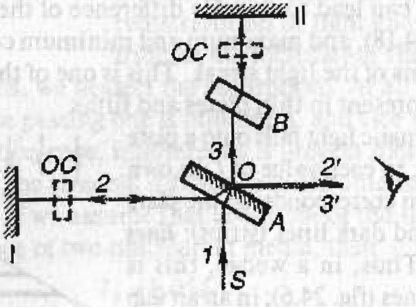


Fig. 24.7

From source *S*, beam *1* of monochromatic light falls at angle 45° onto the flat-parallel plate *A*, whose back surface is translucent, as it is covered with a very thin layer of silver. At point *O*, this beam is split into two beams *2* and *3*, whose intensity is approximately equal. Beam *2* reaches mirror I, is reflected and refracts inside plate *A* and partly leaves the plate (beam *2'*). From point *O*, beam *3* goes to mirror II, is reflected, and returns to plate A, where it is partly reflected (ray *3'*). Beams *2'* and *3'*, getting in the observer's eye, are coherent, and their interference result can be recorded.

---

[1] Strictly speaking, due to multiple reflections, more than two rays can be formed, but their intensity is negligible.

Usually, mirrors I and II are arranged so that from the divergence to meeting points, beams *2* and *3* cover equal paths. To make the optical path of these paths equal, transparent plate *B* is installed across the path of beam 3; it is like plate *A* and installed to compensate two paths passed by beam *2* through plate *A*. In this case, the interference is maximum.

If one mirror is moved over distance $\lambda/4$, then the difference of beam paths will be equal to $\lambda/2$, which corresponds to the minimum, and the interference pattern will move by 0.5 of the fringe[1]. If we move the mirror from the initial position over distance $\lambda/2$, then the optical difference of the path of interfering beams will change by λ, which corresponds to the maximum; we will have interference pattern displacement over the entire fringe. This relationship between mirror displacement and changes in interference pattern makes it possible to measure a wavelength according to mirror displacement and, vice versa, to measure displacement according to wavelength.

A Michelson interferometer is used to measure a refractive index. Across the paths of beams *2* and *3* two identical optical cells *OC* are installed (shown by dashed line in fig. 24.7), one of which is filled with material with refractive index n_1, and the other has a material with refractive index n_2. The optical path difference is

$$\delta = 2ln_1 - 2ln_2 = 2l(n_1 - n_2), \tag{24.23}$$

where l is the length of the single path of the beam inside the medium in the optical cell; since beams pass the optical cell twice, the distance is $2l$. Let us suppose that due to this path difference, the interference pattern moves by k fringes, then

$$\delta = kl. \tag{24.24}$$

Equating (24.23) and (24.24), we get

$$\Delta n = n_1 - n_2 = k\lambda/(2l). \tag{24.25}$$

If we assume that displacement by 0.1 of fringe ($k = 0.1$) can be recorded, then, for instance, at $l = 2.5$ cm, $\lambda = 500$ nm, we have

$$\Delta n = 0.1 \cdot \frac{500 \cdot 10^{-9}}{2 \cdot 0.025} = \frac{5 \cdot 10^{-8}}{5 \cdot 10^{-2}} = 10^{-6}.$$

---

[1] Due to different incidence angle of beams from *S* onto plate *A* or inexact perpendicularity of mirrors I and II, the interference pattern is almost always represented by fringes (fringes of equal inclination or equal thickness respectively). This issue is not considered in detail.

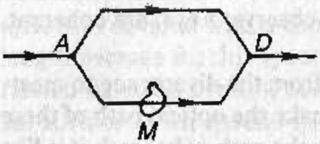


Fig. 24.8

As we can see, an *interference refractom* (one adapted to measure refractive indices i able to detect changes in a refractive index in the sixth decimal place.

Interference refractometerd are used, in particular, for sanitation purposes to det mine the content of harmful gases.

Using the interferometer, Michelson established that the speed of light i independent from the movement of the Earth, which experimental fact serve to develop the special theory of relativity.

A combination of a two-beam interferometer and a microscope is call interference microscope; it is used in biology to measure the refraction inde dry material concentration and thickness of transparent microscopic object

A framework of an interference microscope is shown in fig. 24.8. The bea of light, as in an interferometer, splits at point *A*: one beam passes throu transparent microscopic object *M*, and the other one passes outside At point *D*, the beams meet and interfere; the result of the interference is a cr terion of the parameter being measured.

§ 24.4. HUYGENS–FRESNEL PRINCIPLE

Quantification and explanation of light diffraction can be approximat made using the ***Huygens–Fresnel* principle**.

According to Huygens, each point of a wave surface, which is reache at the moment by the wave, **is a source of elementary secondary waves**; their exterior envelope surface will be a wave surface at the next point of tim (fig. 24.9; S_1 and S_2 are wave surfaces at points of times t_1 and t_2 respective $t_2 > t_1$).

Fresnel supplemented this Huygens' statement having introduced the concept of **coherence** of secondary ***waves and their interference***.

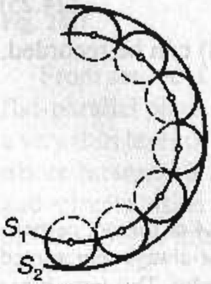


Fig. 24.9

In this generalized form, these ideas are referred to as ***Huygens–Fresnel* principle**.

To determine the result of diffraction at a certain point in space, one should calculate, according to the Huygens–Fresnel principle, the interference of secondary waves which reached this point from the wave surfac For a wave surface of an arbitrary shape, such calculati is rather complicated, but in certain cases (spherical flat wave surface, symmetric location of the point relat to the wave surface and opaque barrier), calculations are relatively simple. The wave surface is divided into separate parts (**Fresnel zones**), arranged in a certain way, which simplifies mathematical operations.

§ 24.5. DIFFRACTION BY A SLIT IN PARALLEL BEAMS

Onto a long narrow slit in a flat opaque barrier MN a plane-parallel beam of monochromatic light falls normally (fig. 24.10; $|AB| = a$ is the slit width; L is a converging lens, in whose focal plane there is screen Sc to observe the diffraction pattern).

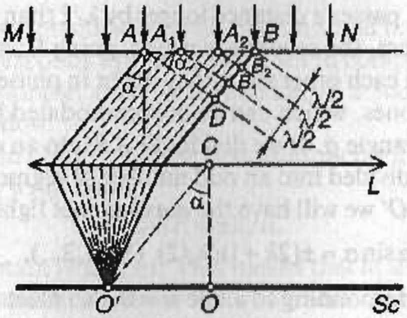


Fig. 24.10

If there was no diffraction, the light beams having passed through the slit would focuse at point O located on the principal optical axis of the lens. Light diffraction by the slit changes the phenomenon very much.

Let us assume that all the rays of the light beam go from the same remote source[1] and, therefore, are coherent. AB is a part of the wave surface each point of which is a center of secondary wave propagation behind the slit in all possible directions. It is impossible to depict all these secondary waves, so in fig. 24.10, we show only secondary waves propagating at angle α towards the direction of incident beam and normal line of the grating. The lens will focus these waves at point O' of the screen, where their interference will be present (the position of point O' is obtained as an intersection between the focal plane and the secondary optic axis CO' of the lens drawn at angle α).

To find out the result of interference of secondary waves, let us do the following generations. Let us draw a perpendicular AD to the direction of the secondary beam. From AD to O', the paths of all secondary waves will be tautochronic, the lens will not add any phase difference between them; so th path difference which resulted between the secondary waves *and* AD, will b the same at point O'.

---

[1] A virtually point source can be placed in the focus of the lens not shown in fig. 24.10, so a parallel beam of coherent waves will propagate from the lens.

Let us split BD into segments equal to $\lambda/2$. In the case shown in fig. 24.1 the following three segments are obtained: $|BB_2| = |B_2B_1| = |B_1D| = \lambda/2$. Fro points B_2 and B_1, let us draw straight lines, which are in parallel to AD, and divide AB into equal Fresnel's zones: $|AA_1| = |AA_2| = |A_2B|$. For any seconda wave, going from any point of a Fresnel zone, in its adjacent zones, we can find corresponding secondary waves so that the difference between their path will be equal to $\lambda/2$.

For example, a secondary wave, going from point A_2 along the chose direction to point O', passes a distance longer by $\lambda/2$ than the wave going from point A_1, etc. Therefore, the secondary waves, going from two adjacent Fresnel zones, will cancel each other since they differ in phase by π.

The number of zones, which can be accommodated by the slits, depend on wavelength λ and angle α. If we divide slit AB into an odd number of Fresnel zones, and BD is divided into an odd number of segments, which are equa to $\lambda/2$, then at point O' we will have the ***maximum*** of light ***intensity***:

$$|BD| = \alpha \sin\alpha = \pm(2k+1)(\lambda/2) \quad (k=1,2...). \qquad (24.2$$

The direction corresponding to angle $\alpha = 0$ also meets the maximum condition since all secondary waves will come into O in equal phase.

If we divide slit AB into an even number of Fresnel zones, then we will hav ***the maximum of*** light ***intensity***:

$$\alpha \sin\alpha = \pm 2k(\lambda/2) = \pm k\lambda \quad (k=1,\ 2...). \qquad (24.27$$

Thus, on screen *Sc*, we will obtain a system of light (maximum) and dark (minimum) fringes, the centers of which correspond to conditions (24.26) or (24.27), located symmetrically on the left and on the right of the centra ($\alpha = 0$), brightest fringe. Intensity I of the other maxima increases in proportion to the distance from the central maximum (fig. 24.11).

If the slit is illuminated with a white light, then on the screen *Sc* [see (24.26), (24.27)], a system of colored fringes is formed, only the central maximum will

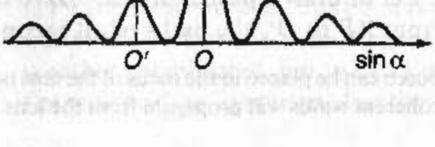


Fig. 24.11

retain the color of the incident light, since at $\alpha = 0$, all light wavelengths are amplified.

Light diffraction, like its interference, is connected with redistribution of electromagnetic wave energy in space. In this sense, a slit in opaque screen is not only a system limiting the energy of a light flux, but a redistributor of this flux in space.

To understand the influence of the ratio between the slit width and wavelength on the probability of observing a diffraction pattern, we will consider the following particular cases.

1. $\lambda \ll \alpha$. Let us represent the formula for maxima in the following form
$$\sin\alpha = \pm(2k+1)\lambda/(2a),$$
we have $\sin\alpha \approx 0$ practically for all maxima, and no diffraction this case. This case corresponds to a rather wide slit in comparison with the wavelength. For example, there is diffraction in a room when light passes through a window.
2. $\alpha \le \lambda$. Based on (24.27), for the first minima limiting the central bright fringe, we can write down:
$$\sin\alpha = \pm\lambda/\alpha.$$

From this, we obtain $|\sin\alpha| \ge 1$. This means that at $\alpha \le \lambda$, instead of a system of minima and maxima, we see the whole screen obscurely lighted. This situation is noted as soon as $\alpha \to \lambda$.

§ 24.6. DIFFRACTION GRATING. DIFFRACTION SPECTRUM

A diffraction grating is an optical device consisting of a large number of parallel, usually equidistant, slits.

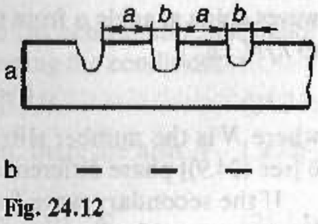


Fig. 24.12

A diffraction grating can be made by making opaque scratches (cuts) on a glass plate. Non-scratched parts, called slits, will pass light; the scratches corresponding to the gaps between the slits disperse light and are opaque. The cross section of such a diffraction grating (a) and its designations (b) are shown in fig. 24.12.

The distance between the centers of adjacent slits is called a grating constant or a period of grating:

$$c = a + b, \tag{24.28}$$

where a is slit width; b is the width of the gap between slits.

If a beam of coherent waves falls onto a grating, the secondary waves goin in all possible directions will interfere, forming a diffraction pattern.

Let a plane-parallel beam of coherent waves fall normally onto the gratin (fig. 24.13). Let us choose a direction of secondary waves at angle α to the normal of the grating. Beams going from the extreme points of two adjacent s have path difference $\delta = |A'B'|$. The same path difference will be seen in secondary waves going from the respectively located pairs of points of adjace slits. If this path difference is divisible by an integer of wavelengths, interf ence will cause *major maxima*, for which the following condition is valid:

$$|A'B'| = k\lambda, \quad \text{or} \quad c\cdot\sin\alpha = \pm k\lambda, \qquad (24.2$$

where $k = 0, 1, 2$ is ***the order of major maxima***. They are located symmetrica relative to the central maximum ($k = 0$, $\alpha = 0$). Equation (24.29) ***is the m formula of diffraction grating.***

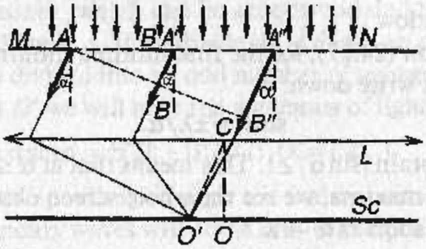


Fig. 24.13

Between the major maxima, minima (additional ones) occur, whose number depends on the number of all the slits of the grating. From this, let us derive a condition for additional minima. Let the path difference of seconda waves going at angle α from the corresponding points of adjacent slits be eq to λ/N, i.e.:

$$\delta = c\cdot\sin\alpha = \lambda/N, \qquad (24.30$$

where N is the number slits in diffraction grating. To this path differen δ [see (24.9)] phase difference $\Delta\varphi = 2\pi/N$ corresponds.

If the secondary wave from the first slit is assumed, at the time point of interference with other waves, to have a phase equal to zero, then the phase of the wave from the second slit is equal to $2\pi/N$, from the third is equal to $4\pi/N$ from the fourth is equal to $6\pi/N$, etc. It is convenient to obtain the result of interference of these waves taking into account the phase difference using a vector diagram: the sum of N equal vectors of an electric (or magnetic) fi intensity, the angle between any adjacent which is $2\pi/N$, is equal to zero. Th means that condition (24.30) corresponds to a minimum. When the path difference of secondary waves from the adjacent slits is $\delta = 2(\lambda/N)$ or the phase difference of $\Delta\varphi = 2(2\pi/N)$, we can also obtain the interference minimum of secondary waves coming from all slits, etc.

As an illustration, fig. 24.14 shows a vector diagram corresponding to a diffraction grating of six slits: E_1, E_2 etc. are the intensity vectors of electric component of electromagnetic waves from the first, second, and other slits.

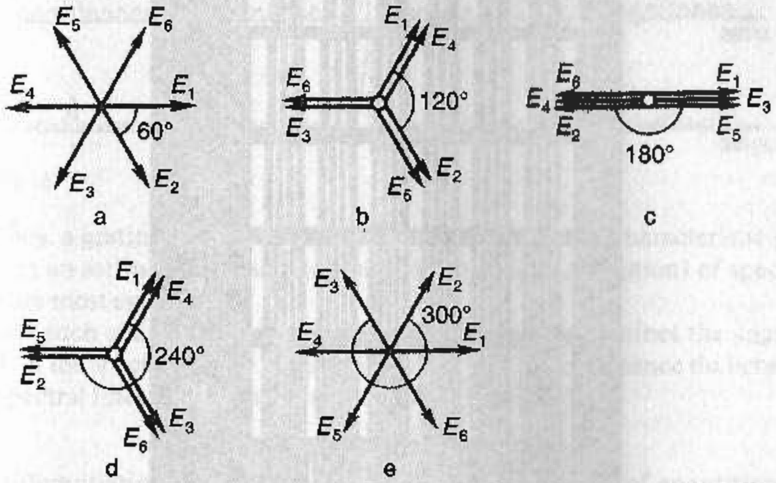


Fig. 24.14

Five additional minima (the sum of vectors is zero) resulting from the interference are observed at a phase difference of waves going from adjacent slits of 60° (a), 120° (b), 180° (c), 240° (d) and 300° (e).

For instance, we can make sure that between the central and first major maximum, there are $N-1$ additional minima meeting the condition:

$$c\sin\alpha = \pm\lambda/N; \pm 2\lambda/N, \dots, \pm(N-1)\lambda/N. \qquad (24.31)$$

Between the first and second major maximum, there are also $N-1$ additional minima meeting the condition:

$$c\sin\alpha = \pm(N+1)\lambda/N; \pm(N+2)\lambda/N \dots (2N-1)\lambda/N, \qquad (24.32)$$

etc. Thus, between any two adjacent major maxima there are $N-1$ additional minima.

When the number of slits is great, individual additional minima are practically indistinguishable, and between the major maxima, the entire space looks dark. The greater the number of slits in the grating, the sharper major maxima are. Fig. 24.15 presents photos of a diffraction pattern obtained from gratin with a varying number N of slits (the grating constant is the same), an fig. 24.16 shows the intensity distribution curve.

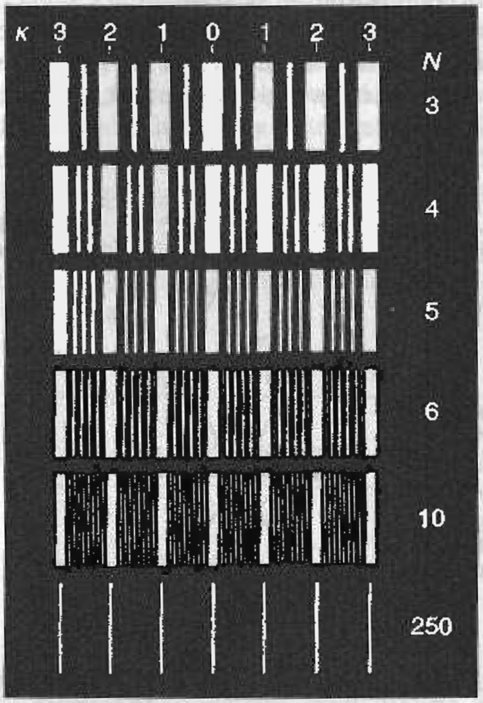


Fig. 24.15

We especially note the role of minima generated by one slit. In the direction, meeting condition (24.27) each slit gives a minimum, so a minimu from one slit will remain the same for the entire grating. If conditions (24.27) and (24.29) are simultaneously valid for a certain direction, no correspon ing major maxima will occur. Usually one tries to use the major maxima, which are placed between the first minima from one slit, that is in the interval:

$$\arcsin(\lambda/\alpha) > \alpha > -\arcsin(\lambda/\alpha). \qquad (24.3$$

When white or other non-monochromatic light falls onto a diffractio grating, each major maximum, except the central one, will be decompos into its constituent colors [see (24.29)]. In this case, k shows ***the order of spectrum***.

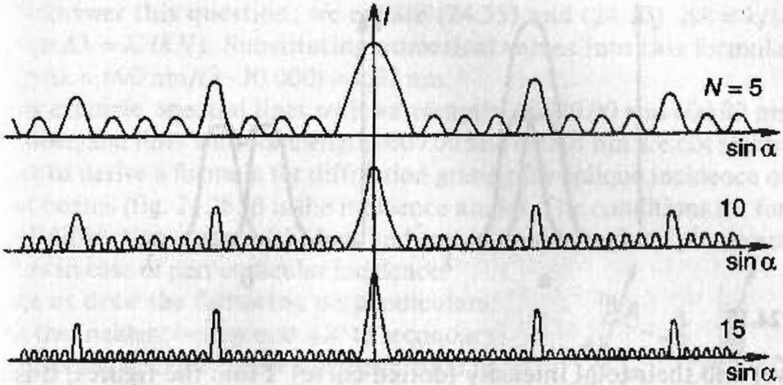


Fig. 24.16

Thus, a grating, like a slit, is a spectral instrument, so characteristic permitting an estimate of possibility of distinguishing (resolution) of spectral lines are most essential.

One such characteristic — ***angular dispersion*** — determines the angular width of the spectrum. It is numerically equal to angular distance dα between two spectral lines, whose wavelengths differ by one ($d\lambda = 1$):

$$D = d\alpha / d\lambda.$$

Differentiating (24.29) and using only positive values of quantities, we obtain:

$$c \cdot \cos\alpha \, d\alpha = k \cdot d\lambda.$$

From the last two equations, we have:

$$D = k/(c \cdot \cos\alpha). \qquad (24.34)$$

Since gratings mostly have small angles, $\cos\alpha \approx 1$. The greater order k of the spectrum and the smaller constant c of a diffraction grating, the higher angular scattering.

The ability to distinguish close spectral lines depends not only on spectrum width or angular dispersion, but also on the width of the lines that can overlap.

It is generally accepted that if there is a region between two maxima of equal intensity where the intensity is 80% of the maximum, then the spectral lines corresponding to the maxima are already resolvable. In this case, according to J.W. Rayleigh, the maximum of one line coincides with the nearest minimum of the other, which is considered the criterion of resolution. Fig. 24.17 shows dependencies I of separate lines on the wavelength (solid

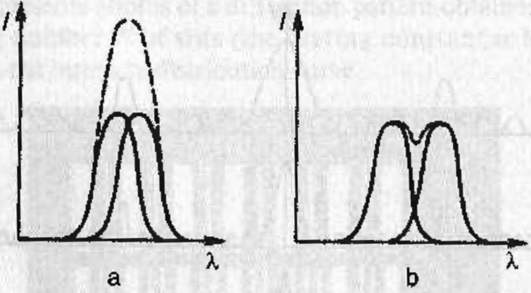


Fig. 24.17

curve) and their total intensity (dotted curve). From the figures, it is easy to see where two lines are not resolvable (a) or maximally resolvable (b), wh the maximum of one line coincides with the nearest minimum of the other.

Resolution of spectral lines is quantified by ***resolution capability***, which is equal to the ratio between the wavelength to the smallest interval of wavelengths that can still be resolved:

$$R = \lambda/\Delta\lambda. \qquad (24.35)$$

If there are two close waves $\lambda_1 \approx \lambda_2$, $\Delta\lambda = \lambda_1 - \lambda_2$, then formula (24.35) can be written in the following approximate form:

$$R = \lambda_1/(\lambda_1 - \lambda_2), \quad \text{or} \quad R = \lambda_2/(\lambda_1 - \lambda_2). \qquad (24.36)$$

For the first wave, the major maximum condition is $c \cdot \sin\alpha = k \cdot \lambda_1$.

The closest minimum for the second wave coincides with it, the condition for the former is:

$$c \cdot \sin\alpha = k\lambda_2 + \lambda_2/N.$$

Equating the right-hand members of the last two equations, we have:

$$k \cdot \lambda_1 = k \cdot \lambda_2 + \lambda_2/N, \quad k(\lambda_1 - \lambda_2) = \lambda_2/N,$$

whence, [see (24.36)]:

$$R = k \cdot N. \qquad (24.37)$$

So then, the higher the resolution of a diffraction grating, the greater the order k of the spectrum and number N of lines.

Example

In a spectrum obtained with a grating with number $N = 10\,000$ of slits there are two lines near wavelength $\lambda = 600$ nm. At what smallest wavelength difference do these lines differ in the spectrum of the third order ($k = 3$)?

To answer this question, we equate (24.35) and (24.37): $\Delta\lambda = \lambda/(kN)$, whence $\Delta\lambda = \lambda/(kN)$. Substituting numerical values into this formula, we obtain $\Delta\lambda = 600\ \text{nm}/(3 \cdot 10\,000) = 0.02\ \text{nm}$.

For example, spectral lines with wavelengths of 600.00 and 600.02 nm are resolvable, and lines with wavelengths 600.00 and 600.01 nm are not resolvable.

Let us derive a formula for diffraction grating for oblique incidence of coherent beams (fig. 24.18, β is the incidence angle). The conditions for formation of diffraction pattern (the lens and screen are in the focal plane) are the same as in case of perpendicular incidence.

Let us draw the following perpendiculars: $A'B$ to the incident beams and AB' to secondary waves passing at angle α to the perpendicular to the grating plane.

From fig. 24.18, we can see that relatively to position $A'B$, the beams have equal phases and from AB' and further, the phase difference is the same. Hence, path difference:

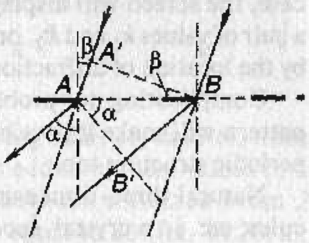


Fig. 24.18

$$\delta = |AA'| - |BB'|. \tag{24.38}$$

From $\Delta AA'B$, we have $|AA'| = |AB| \cdot \sin\beta = c \cdot \sin\beta$. Considerring $\Delta BB'A$, we obtain $|BB'| = |AB| \cdot \sin\alpha = c \cdot \sin\alpha$. Substituting the expressions for $|AA'|$ and $|BB'|$ into (24.38) and taking into account the condition for the major maxima, we have:

$$c(\sin\beta - \sin\alpha) = \pm k\lambda. \tag{24.39}$$

The central major maximum corresponds to the direction of incident beams ($\alpha = \beta$).

Along with transparent diffraction gratings, there are reflecting gratings with ruling on their metal surface. In this case, we can see interference in the reflected light. Reflecting diffraction gratings made on a concave surface can form a diffraction pattern without a lens.

In modern diffraction gratings, the maximum number of grooves is more than 2,000 per 1 mm, and the maximum size of the ruled surface is 300×300 mm, which gives a quantity for N of about a million.

§ 24.7. FUNDAMENTALS OF X-RAY CRYSTALLOGRAPHY

The basic formula (24.29) for a diffraction grating can be used not only to determine the wavelength, but also to solve the inverse problem: finding the constant of diffraction grating by the wavelength known. Such a modest problem, as applied to a common diffraction grating, leads to a practical important problem: measurement of crystal lattice parameters by X-ray diffraction, which is the content of X-ray crystallography.

Let two diffraction gratings with perpendicular grooves be combined The gratings meet the conditions of the major maxima:

$$c_1\cdot\sin\alpha_1=\pm k_1\lambda,\quad c_2\cdot\sin\alpha_2=\pm k_2\cdot\lambda.\qquad (24.4$$

Angles α_1 and α_2 are counted in mutually perpendicular directions. In this case, the screen will display a system of spots, each of which corresponds to a pair of values k_1 and k_2, or α_1 and α_2. Therefore, in this, we can find c_1 and c_2 by the location of diffraction spots.

Complicating our problem, it would belogical to think that a diffraction pattern will make it possible to measure parameters of a three-dimensional periodic structure too.

Natural three-dimensional periodic structures are crystals, large molecules, etc. In a crystal, secondary waves occur because of interaction of primary beams with the electrons of atoms.

For a clear observation of a diffraction pattern, a certain relationship between the wavelength and the parameter of a periodic structure should be valid (see § 24.5). The optimal conditions correspond to an approximately equal order of these values. Taking in account the fact that in a crystal, the distance between the scattering centers (atoms) ($\sim 10^{-10}$ m) is approximately equal to the wavelength of X-rays, we can consider that for these beams, a crystal is a three-dimensional diffraction grating.

In fig. 24.19, the dotted line shows two adjacent crystallographic planes Simplifying the process, the interaction of X-rays with atoms and generation of secondary waves can be considered as reflection from the planes.

Let X-rays *1* and *2* fall onto a crystal at angle θ; *1'* and *2'* are reflected (secondary) beams, *CE* and *CF* are perpendiculars to the incident and reflected beams respectively. The path difference of reflected beams *1'* and *2'* is:

$$\delta=|DE|+|DF|=2l\sin\theta,\qquad (24.41)$$

where *l* is interplanar spacing.

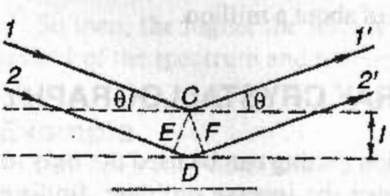


Fig. 24.19

Maxima of interference occur upon reflection when the path difference is equal to an integer number of wavelengths:

$$2l\sin\theta=k\lambda\quad(k=1,2,3...).\qquad(24.42)$$

This is ***Wulff–Bragg's formula.***

When a monochromatic X-ray falls onto the crystal at different angles

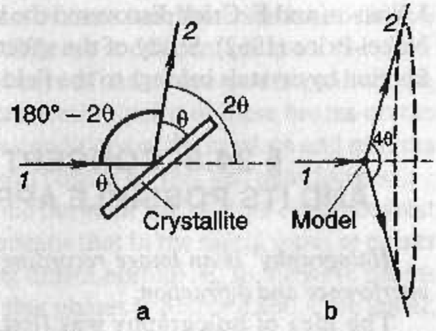


Fig. 24.20

the strongest reflection (maximum) will be at angles meeting condition (24.42). When we observe a sliding X-ray beam at a certain sliding angle with a continuous spectrum, the maximum diffraction will be valid for wavelengths satisfying Wulff–Bragg's condition.

P. Debye and P. Scherrer suggested a method of X-ray crystallography based on diffraction of monochromatic X-rays inside polycrystalline bodies (they are usually compressed powders). Among many crystallites, there are always those, for which l, θ and k are equal; in addition, these values correspond to Wulff–Bragg's formula. The angle between reflected beam *2* (the maximum) and the incident X-ray L will be equal to 2θ (fig. 24.20, a). Since condition (24.42) is the same for many crystals with different orientation, the X-rays diffracted form a cone in space; the apex of the cone lies in the object being studied, and the opening angle is equal to 4θ (fig. 24.20, b). Another set of values l, θ and k, meeting the condition (24.42), corresponds to another cone. On a photographic film, X-rays will form a radiograph (Debye powder diagram) in the form of circumference (fig. 24.21) or arc.

X-ray diffraction is also observed when X-rays are scattered by amorphous bodies, liquids, and gases. In this case, the radiograph has wide, blurred rings.

Nowadays, X-ray crystallography of biological molecules and systems is widely used: fig. 24.22 shows radiographs of proteins. Using this method,

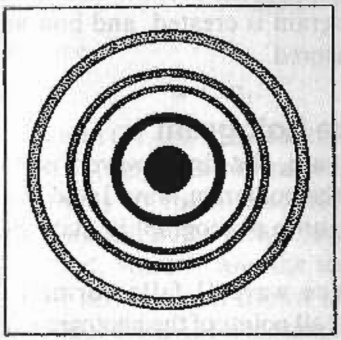


Fig. 24.21

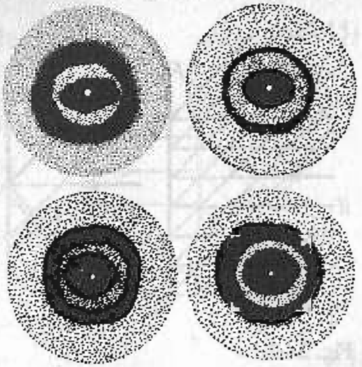


Fig. 24.22

J. Watson and F. Crick discovered the structure of DNA and were awarded th Nobel Price (1962). Study of the spectral composition of X-rays by their dif fraction by crystals belongs to the field of X-ray spectroscopy.

§ 24.8. CONCEPT OF HOLOGRAPHY AND ITS POSSIBLE APPLICATION IN MEDICINE

Holography*[1] *is an image recording and restoration method based on wa interference and diffraction.

The idea of holography was first voiced by D. Gabor in 1948; nev theless, its practical use became possible after invention of lasers.

It is not out of place to begin a discussion of of holography with compar it with photography. When a photograph is taken, the intensity of light wa reflected by an object is recorded on the photographic film. In this cas the image is a cluster of dark and light dots. The phases of scattered waves a not recorded, thus, much information about the object is lost.

Holography makes it possible to record and reproduce more detailed in mation about an object, taking into account the amplitudes and phases of th waves scattered by the object. Phase recording is possible due to wave inter rence. For this purpose, two coherent waves are sent onto the light recordi surface: the reference wave going from a light source or a mirror directly (mi rors being auxiliary devices), and the signal wave that is generated for scatt ing (reflection) of a part of the reference wave by the object and contains rel vant information about it.

An interference pattern formed by superposition of the signal and referen waves, and recorded on a photosensitive plate, is called a hologram. **To resto the image the hologram is illuminated by the same reference wave.**

Let us show with certain examp how a hologram is created, and how an image is restored.

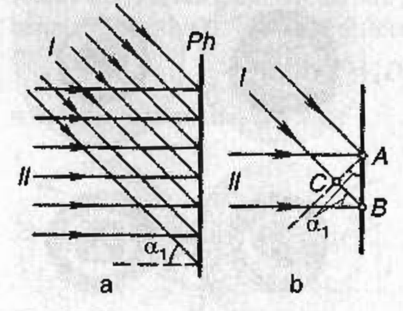


Fig. 24.23

Flat wave hologram

In this case, flat signal wave I is re corded on the hologram, wave Incide angle is α_1 onto photographic plat Ph (fig. 24.23).

Reference wave II falls norma therefore, at all points of the photogra plate, at the same time point, the phase is the same too. Due to its inclined incidence the phases of the signal wave are different at different points of the light-sensitive layer. Whence, between the rays of reference and signal waves, the phase difference depends on the intersection point of these beams on the photographic plate and, according to the conditions of the maxima and minima of interference, the resulting hologram will consist of dark and light fringes.

---

[1] Holography (Greek) means a method of complete recording.

Let AB (fig. 24.23, b) be the distance between the centers of the nearest dark or light interference fringes. This means that in the signal wave, at points A and B, the phases differ by 2π. Having drawn normal AC to the wave beams (the wave front), it is not difficult to see that phases at points A and C are equal. The phase difference at points B and C by 2π means that $|BC| = \lambda$. From rectangular ΔABC, we have

$$|AB| = |BC|/\sin\alpha_1 = \lambda/\sin\alpha_1. \tag{24.43}$$

So, in this example, the hologram is like a diffraction grating since the areas of amplified (maxima) and weakened (minima) oscillations are registered on a photosensitive surface, distance AB between which is determined by formula (24.43).

Since the signal wave is formed when a part of the reference wave is reflected from the object, it is clear that in this case the object is a plane mirror or prism, that is a device that converts a plane reference wave into a plane signal wave (technical details are not shown in fig. 24.23, a).

By directing reference wave I onto the hologram (fig. 24.24), we perform diffraction (see § 24.6). According to (24.29), the first major maxima ($k = 1$) correspond to the directions

$$\sin\alpha = \pm\lambda/c. \tag{24.44}$$

Substituting AB from (24.43) instead of c, we have

$$\sin\alpha = \pm\lambda\sin\alpha_1 = \pm\sin\alpha_1 \tag{24.45}$$

whence

$$\alpha = \pm\alpha_1. \tag{24.46}$$

From (24.46), it is clear that the direction of wave I' (fig. 24.24) diffracted at angle α_1 corresponds to the signal wave: in this way, the wave reflected (scattered) by an object is restored. Wave I'' and the waves of the other major maxima (not shown in the figure) also reproduce the information recorded in the hologram.

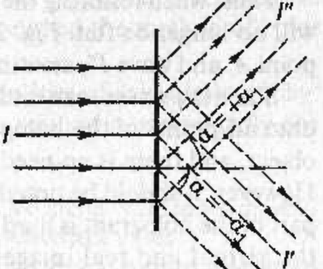


Fig. 24.24

Hologram of a point

One part of the reference wave *II* falls onto point object *A* (fig. 24.25, a) and is scattered from it in the form of spherical signal wave *I*, the other part is directed by flat mirror *M* to photographic plate *Ph* where these waves interfere. The source of radiation is laser *L*. Fig. 24.25, b shows sketchily the hologram obtained.

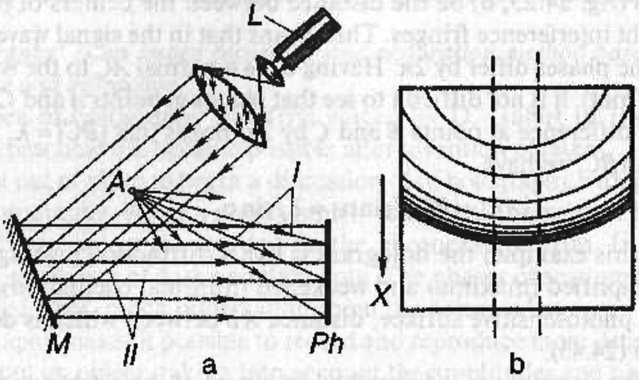


Fig. 24.25

Although in this example the signal wave is spherical, with a certain approximation, we can apply formula (24.45) and notice that when angle α_1 increases (see fig. 24.23, a), distance *AB* between adjacent fringes reduc On the hologram, the lower arcs (fig. 24.25, b) are located closer one to anoth

If we cut out a narrow fring from the hologram shown by dotted lines in fig. 24.25, b, it will be like a narrow diffraction grating, whose constant decreases in the direction of *X*-axis. On such a grating, deflection of seconda waves corresponding to the first major maximum increases as coordinate *x* of slit increases [see (24.41)]: *c* decreases, and $|\sin\alpha|$ increases.

Thus, when restoring the image of flat reference wave, the diffracted wav will no longer be flat. Fig. 24.26 shows wave *I'*, forming virtual image *A'* of point *A*, and wave *I''*, creating a real image of *A''*.

Since the waves scattered by the object together with the reference wave fall onto all points of the hologram, all its parts contain information about the object, and there is no need to use the entire hologram to restore the imag However, it should be noted that the restored image is the worse, the small part of the hologram is used for this purpose. From fig. 24.26, we can see that the virtual and real images are formed even if restoration is carried out, for example, using the lower half of the hologram (dashed lines); moreov the image is formed by a smaller number of beams.

Any object is a set of points, so a reasoning given for one point, can be generalized to the whole hologram of any object. Holographic images are threedimensional and their visual perception is no different from the perception of the corresponding objects[1]: clear vision of different points of the image is implemented by means of adaptation of the eye (see § 26.4); upon changing the angle of view the perspective changes too, certain details of the image can obscure others.

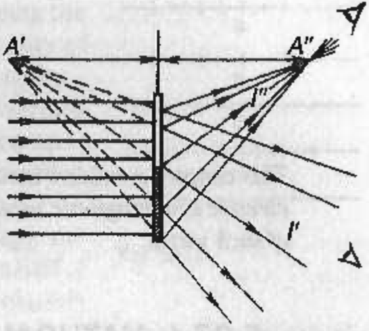


Fig. 24.26

When restoring an image, the reference wavelength can be changed. For example, a hologram formed by invisible electromagnetic waves (ultraviolet, infrared, and X-rays) can be restored by visible light. Since conditions of reflection and absorption of electromagnetic waves by bodies depend, in particular, on the wavelength, this feature of holography makes it possible to use it as a method of *diagnostic imaging*, or *introscopy*[2].

Particularly interesting and important prospects open in connection with ultrasonic holography. Having obtained a hologram with ultrasonic mechanical waves it is possible to restore it with visible light. In the future, ultrasonic holography can be used in medicine to examine the viscera of a person to make his or her diagnosis, to determine the sex of fetus in utero, etc. Taking into account the high informative value of this method and much less harm of ultrasound in comparison with X-rays, we can expect that in future, ultrasonic holographic introscopy will replace traditional X-ray diagnostics.

Another biomedical application of holography is connected with the holographic microscope. This instrument is based on the imaging an object to be enlarged if a hologram recorded with a flat reference wave is illuminated with a divergent spherical wave.

A Soviet physicist, who is a winner of the Lenin Prize, Yu.N. Denisyuk, contributed to holography development. He developed the method of color holography.

Nowadays, it is difficult to evaluate all possibilities of using holography: cinema, television, data storage devices, etc. This method is undoubtedly one of the greatest inventions of our time.

---

[1] Some difference is due to the monochromaticity of the image. Such monochromaticity is inevitable in recording and restoring with monochromatic waves.

[2] *Intro* is the Latin for inside, and *scopeo* is the Greek for view. It is visual observation of objects, phenomena, and processes in optically opaque bodies and media as well as in conditions of poor visibility.

Chapter 25

Light polarizatio

This chapter considers methods of obtaining an ordered orientation of th electric and magnetic vectors of light waves as well as certain properti of such waves.

§ 25.1. NATURAL AND POLARIZED LIGHT. MALUS'S LAW

An electromagnetic wave in which vectors $\vec{E}$ and hence, $\vec{H}$ are in qu definite planes is called a plane polarized wave.

The plane passing through the electric vector $\bar{E}$ in the direction of an electromagnetic wave propagation is the plane of polarization.

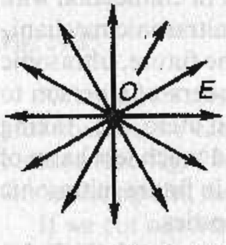


Fig. 25.1

A plane-polarized wave is radiated by a separate atom. In the natural light going from the sun, in an incandescent filament of a lamp, a gas-discharge tub, or flame and the like, disordered radiations of mult tude of randomly oriented atoms are summed up, tha is why vectors $\vec{E}$ are not coplanar[1]. Such light can b considered as superposition of plane-polarized wave with chaotic orientation of oscillation planes, whos electric vectors are oriented in all possible directio which are perpendicular to the beam. Fig. 25.1 shows the cross sectio of beam O and projections of vectors $\vec{E}$ onto the plane perpendicular to the bea at a certain point of time.

If we select any two mutually perpendicular planes, passing through a ray of natural ligh, and project vectors $\vec{E}$ onto these planes, on average, these p jections will be equal. Therefore, it is convenient to depict a ray of natural li as a straight line on which an equal number of these and other projectio in the form of dashes and dots are displayed (fig. 25.2, a). Thus, a straight ne with dashes (fig. 25.2, b) or dots (fig. 25.2, c) represents a ray of pla polarized light.

A ray of light consisting of natural and polarized components described as partially polarized is conventionally shown in fig. 25.2, d, e, and the ratio

---

[1] In fact, many sources emit partially polarized light.

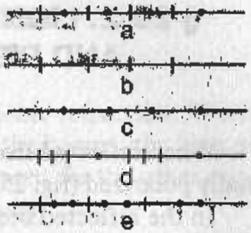


Fig. 25.2

between the number of dashes and dots reflects the **degree of polarization**, i.e. the portion of intensity of the polarized component in the total light intensity.

A device transforming natural light into polarized light is called a **polarizer**. It transmits only a component of vector $\vec{E}$ onto a certain plane, which is called the main **polarizer plane**.

In this case, polarized light[1] passes through the polarizer; the intensity of this light is equal to a half of the intensity of the incident light. When the polarizer rotates around a ray of natural light, the plane of oscillations of the plane polarized light emitted by the device, rotates too, but its intensity does not change. A polarizer can be used to analyze polarized light; in this case, it is called an analyzer.

If plane-polarized light with an amplitude of electric vector $\vec{E}_0$ falls onto an analyzer, it will transmit only the component of vector $\vec{E}_0$; the component is equal to:

$$E = E_0 \cos\varphi, \tag{25.1}$$

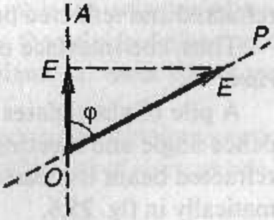


Fig. 25.3

where φ is the angle between the principal planes of polarizer P and analyzer A (fig. 25.3).

Since light intensity is proportional to the square of the oscillation amplitude [see (7.16)], then from (25.1) we obtain:

$$I = I_0 \cos^2 \varphi, \tag{25.2}$$

where I_0 is the intensity of plane-polarized light falling onto the analyzer; I is the intensity of light emitted by the analyzer. Equation (25.2) expresses **Malus's law.**

As we can see from Malus's law, when the analyzer rotates around the beam of incident plane-polarized light, the intensity of emitted light changes from zero to I_0. If the intensity of emitted light does not change when the analyzer rotates around the incident beam as an axis of rotation, the light can be natural[2]; if its intensity changes according to law (25.2), the incident light is plane-polarized one.

The term "light polarization" has two meanings. First, it means the property of light to be an ordered orientation of electric and magnetic vectors in space and time. Second, light polarization is the process of obtaining polarized light.

---

[1] Possible absorption of light by the polarizer is not taken into account here.

[2] At all that, circular polarization is also possible, but it is not considered here.

§ 25.2. LIGHT POLARIZATION FOR REFLECTION AND REFRACTION AT THE INTERFACE OF TWO DIELECTRICS

When reflected from the interface of two dielectrics, natural light is pa tially polarized (fig. 25.4).

In the reflected beam oscillations prevail, which are perpendicular to th plane of incidence; in a refracted beam — those, which are parallel to th plane. If the incidence angle meets the following condition

$$\tan i_B = n, \qquad (25.3)$$

then the reflected beam *is completely plane-polarized* (fig. 25.5). Formula (25.3) expresses the *Brewster law*. Here, the incidence angle i_B is the Brewster angl or *the angle of complete polarization*; n is the relative refractive index of tw media.

The refracted beam (if the Brewster law is valid) is partially polariz in this case the degree of its polarization is the highest.

Using (25.3) the law of refraction additionally, it is not difficult to sho that at complete polarization of the reflected light, the angle between th refracted and reflected beams is 90°.

Thus, the interface of two dielectrics or dielectric and vacuum is a ***pola*** ***rizer.***

A pile of glass plates is also used as a polarizer. Irrespective of the inci dence angle and meeting the Brewster law, the degree of polarization of th refracted beam increases as it passes through the plates; this is shown sch matically in fig. 25.6.

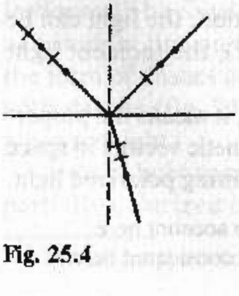


Fig. 25.4

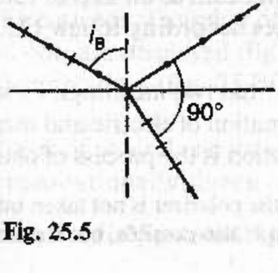


Fig. 25.5

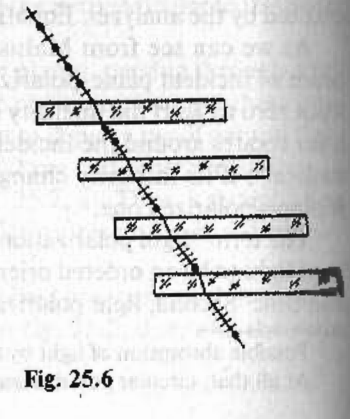


Fig. 25.6

§ 25.3. LIGHT POLARIZATION BY DOUBLE REFRACTION

Some transparent crystals have the property of double refraction: When light hits the crystal, the beam bifurcates. For one of the beams, the laws of refraction are valid, therefore, this beam is called *ordinary*, for the other one, the laws are not valid, and the beam is called *extraordinary*.

At normal incidence of light onto a crystal surface, double refraction is shown in fig. 25.7: the ordinary (o) beam, according to the law of refraction, is not refracted; the extraordinary (e) beam is refracted.

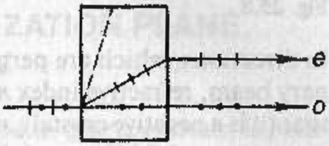


Fig. 25.7

The directions, in which there is no double refraction, and both the ordinary and extraordinary beams have equal speed, are called *crystal optical axes* (dotted line in fig. 25.7). If there is one such direction, such crystals are called uniaxial (exactly this kind is considered in this section). These are Iceland spar (a variety of calcium carbonate $CaCO_3$, crystals of hexagonal structure), quartz, tourmaline (complex aluminosilicate, crystals of trigonal structure), etc. The plane, passing through the optical axis and the incident beam, is the ***principal*** plane. Oscillations of the ordinary beam are perpendicular to the principal plane, and those of the extraordinary beam are in the principal plane, i.e. these beams are polarized in mutually perpendicular planes.

Double refraction is due to specifics of propagation of electromagnetic waves in anisotropic media: amplitudes of forced oscillations of electrons depend on directions of these oscillations.

In crystals, the passage of ordinary and extraordinary beams can be represented as wave surfaces. Let us assume that inside the crystal, there was a light flash, and two waves, ordinary and extraordinary ones, propagate in all directions. At a certain point of time, their wave surfaces will take the position shown in fig. 25.8 (case a is for positive crystals; case b is for negative ones). The spheres correspond to ordinary waves with equal speed υ_o in all directions; the ellipsoids correspond to extraordinary ones, whose speed υ_e depends on direction. Along optical axes OO', the speed of ordinary and extraordinary waves is equal and is equal to:

$$\upsilon_o = c/n_o, \tag{25.4}$$

where n_o is the refractive index of an ordinary beam, which varies for different crystals.

For positive crystals, $\upsilon_e \leq \upsilon_o$, for negative crystals $\upsilon_e \geq \upsilon_o$. The ordinary and extraordinary waves show the highest difference of their velocities

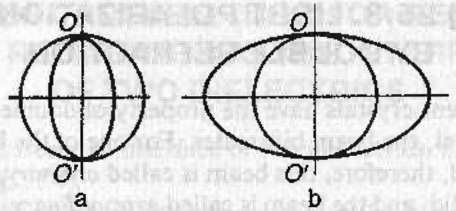


Fig. 25.8

in directions, which are perpendicular to the optical axis; for the extraordinary beam, refractive index n_e is introduced for these directions. For Iceland spar (it is a negative crystal), $n_o = 1.6584$, $n_e = 1.4864$; for quartz (it is a posit crystal) $n_o = 1.5442$, $n_e = 1.5533$ (the data are given for sodium yellow light $\lambda = 589.3$ nm).

Double refraction crystals are not directly used as polarizers since bea of ordinary and extraordinary rays are not separated sufficiently or even ov lap. However, special polarization prisms are made from these crystals.

Let us consider the most commonly used prism proposed by W. Nicol (a ***Nicol prism***).

A Nicol prism is one made from Iceland spar, cut diagonally and glued with Canada balsam[1], *C* (fig. 25.9). For it, $n = 1.550$; this value is between the refractive indices of ordinary and extraordinary beams inside Iceland spa This permits having chosen the appropriate angles of the prism, to ensu complete reflection of the ordinary (*o*) beam at the interface with the Can balsam. In this case, the reflected beam is absorbed by the blackened lo surface or leaves the crystal. The extraordinary beam (*e*) leaves the Nicol prism parallel to the lower surface.

Another operational principle is used in polarizers manufactured from tourmaline, herapathite (iodinoquinine sulphate) and some other crystals, which along with their double refraction, can absorb a certain beam much strong than they do other beams (dichroism). Thus, in a plate of tourmaline about 1 mm thick, an ordinary beam is absorbe practically completely, and the light, which has left the crystal, is plane-polarized.

On celluloid film, very fine crystals of herapathite cover sizeable areas. An electric field orientates them. Such devices (polarizing filters) can operate as polarizers (analyzers).

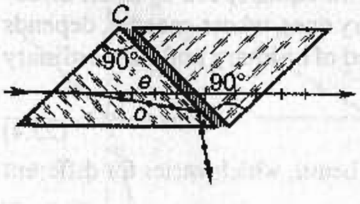


Fig. 25.9

---

[1] This is a resinous material — turpentine — obtained from balsam fir.

In comparison with a Nicol prism, the main disadvantage of tourmaline and polarizing filters is their poor spectral properties. Having left these polarization devices, white light transforms into colored light, while a Nicol prism is transparent in the visible spectrum.

The advantage of polarizing filters is a large surface area that makes it possible to use wide light beams.

§ 25.4. ROTATION OF POLARIZATION PLANE. POLARIMETRY

Rotation of polarization plane, first discovered in quartz crystals, is polarization twist of the plane-polarized light when it passes through a substance. ***Materials with this property are called optically active.***

Let monochromatic light fall from source *S* onto a system of the polarizer *P*–analyzer *A* (fig. 25.10) which are placed crosswise, that is their principal axes are mutually perpendicular. In this case, the light will not reach the observer since the analyzer does not transmit plane-polarized light in accordance with Malus's law ($\varphi = 90°$).

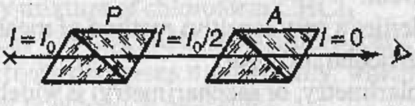


Fig. 25.10

If we place a quartz plate between the polarizer and analyzer so that light propagates along its optical axis, then in a general case, the light will reach the observer. If we rotate the analyzer through a certain angle, we can attain darkening again. This indicates that the quartz plate made the polarization plane rotate through an angle corresponding to the analyzer twist for attaining darkening.

In the experiment, changing the wavelengths, it is possible to observe scattering of rotation of the polarization plane (rotational dispersion), i.e. a dependence of the rotation angle on the wavelength. A quartz plate 1 mm thick rotates the polarization plane approximately through the following angles (Table 25.1).

Table 25.1

Color of a light beam	α, deg.
Red	15
Yellow	21
Violet	51

For a definite wavelength, angle α of polarization plane twist is propo tional to distance l covered by light inside an optically active material:

$$\alpha = \alpha_0 l, \qquad (25.5)$$

where α_0 is the proportionality coefficient or ***specific rotation*** (rotation cap bility), deg/mm.

There are two modifications of quartz, each of which rotates the pla of polarization in a certain direction: clockwise for right-handed (positi quartz, counter-clockwise[1] for left-handed (negative) quartz. The speci rotation is the same in both cases.

Many non-crystalline solids are also optically active: pure liquids (like turpentine), solutions of optically active materials in inactive solvents (water sug solution), certain gases, and vapors (camphor vapors).

For solutions, the following quantitative law was determined:

$$\alpha = [\alpha_0]Cl, \qquad (25.6)$$

where C is concentration of an optically active substance; l is the solution lay thickness; $[\alpha_0]$ is specific rotation that is approximately inversely proportio to the square of the wavelength and depends on the actual temperature and properties of the solvent.

Ratio (25.6) underlies a very sensitive method of measuring the concent tion of solutes, in particular, of sugar.

This method (polarimetry, or saccharimetry) is widely used in medicine to determine the concentration of sugar in urine, in biophysical studies, as well as in the food industry. The corresponding measuring instruments are called polarimeters or saccharimeters.

A polarimeter makes it possible to measure not only the concentration, but also specific rotation. Using different light filters, it can find dependen of specific rotation on the wavelength (optical rotatory dispersion). Nowada special devices (spectropolarimeters) are used for this purpose.

Rotation of the polarization plane by solutions is due to interaction of an electromagnetic wave with asymmetric molecules of an optically active solut Such molecules do not have mirror symmetry, i.e. when they are reflected in a mirror, the result is different shape. A left-handed molecule is a mirr image of a right-handed one. Molecules with the same chemical formula but different structure rotate the plane of polarization in different directions.

It is characteristic that all most important biological molecules (protei nucleic acids, polysaccharides, etc.) are asymmetric and can be represented in pairs as antipodes, each of which is the mirror image of the other. Howev only one optical antipode is usually present in materials of biological rather than synthetic origin. For example, sugar obtained conventionally is dextrorotatory, but when it is synthesized by chemical methods, the final mixture contains an equal number of right and left molecules. Such a mixture is called *racemic*; it does not rotate the plane of polarization, as the action of some molecules cancels the others. If bacteria that feed on sugar are placed into a synthetically produced sugar solution, they will assimilate only molecules of dextrorotatory sugar.

---

[1] The direction of rotation is set relative to the observer facing the light ray.

A racemic mixture is a less ordered system and has greater entropy than a similar set of molecules all of the same type. This thermodynamic difference between synthetic and natural materials can be an illustration of the physical meaning of entropy in biological systems.

Polarimetry is used not only to determine the concentration of solutions, but also as a method of studying structural transformations, in particular in molecular biophysics. As an example, fig. 25.11 shows the curve of specific rotation change $[\alpha_0]$ in a polypeptide depending on the composition of a solvent, which is a binary mixture of chloroform $CHCl_3$ and dichloroacetic acid $CHCl_2COOH$. At dichloroacetic acid concentration of 80%, its optical activity decreases dramatically, which indicates a change in the conformation of polypeptide molecules.

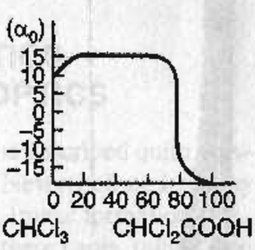


Fig. 25.11

§ 25.5. STUDY OF BIOLOGICAL TISSUES IN POLARIZED LIGHT

Looking at transparent biological objects through a microscope it is difficult to identify different structures, so you have to use certain special techniques, including ***polarization microscopy.***

A polarization microscope is like a conventional biological microscope, but has a polarizer in front of the condenser and an analyzer in the tube between the lens and the eyepiece. The object table can rotate around the optical axis of the microscope. Thus the object is illuminated with polarized rays and examined through the analyzer.

If we cross a polarizer and an analyzer, the field of view will be dark, it will remain the same when we place isotropic transparent bodies onto the object table. Anisotropic objects change our field of view depending on the effect, which they will exert on the direction of the plane of oscillation of polarized light.

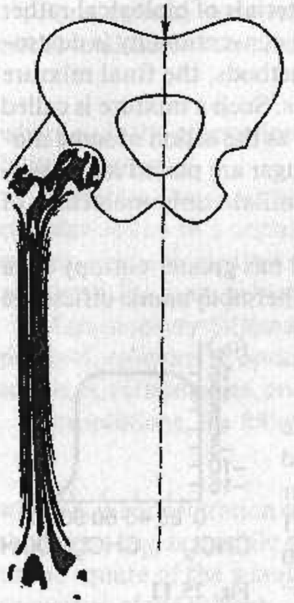


Fig. 25.12

Since certain tissues like muscle, bone, nerv have optical anisotropy, polarization microsc of biological objects is possible. When the polarizer and the analyzer are crossed, only t fibers, whose anisotropy changes the polar light, will be visible.

Polarized light can be used in model conditions to estimate mechanical stress arising in osseous tissue.

This method is based on the phenomenon of *photoelasticity*, which consists in optical anisotropy in initially isotropic solids because of mechanical loads.

A transparent isotropic material such as acrylic plastic is used to create a flat bone model. Through crossed polarizing filters, this model is not visible, as it looks dark. A force applied induces anisotropy of acrylic plastic which bec visible by the distinctive pattern of stripes and spots (fig. 25.12). By this picture, as well as by its change upon increasing or decreasing load, it is possible to make conclusions about mechanical stresses inside the model, and therefore, inside the real object.

Chapter 26

Geometric optics

Geometrical (ray) optics is a section of physics, in which the laws of light propagation are studied based on the following idea: a beam of light is a line, along which the energy of a light wave propagates.

This chapter deals with laws of geometrical optics applied to specific optical systems. Along with this, problems of physics of the eye are represented.

§ 26.1. GEOMETRICAL OPTICS AS LIMITING CASE OF WAVE OPTICS

In the previous chapters, we considered the phenomena described quite satisfactorily in the framework of the wave nature of light. Nevertheless, in many practical applications, such as formation of a light beam, image formation, etc., the wave nature of light can be inessential. Moreover, in these cases, taking into account interference, diffraction and polarization will only complicate obtaining a result. To solve such problems, the laws of geometrical optics are used.

Geometrical optics is the limiting case of wave optics when the wavelength tends to zero.

This can be explained with an example of diffraction grating. From (24.26), when $\lambda \to 0$, it follows $\alpha \to 0$, i.e. a parallel beam of light is focused at point O of the focal plane (see fig. 24.10) which is a common case for a lens.

To find out the limiting possibilities of optical systems, it is again necessary to take the wave nature of light into account. Therefore, this chapter deals with interference and diffraction partially.

Geometrical optics is an example of a theory that with a small number of basic concepts and laws, yields many practically important results. In the theory of optical devices, it is still of great importance.

§ 26.2. LENS ABERRATION

Only spherical lenses with convex or concave spherical surfaces are considered in this course. The focal distance of such a lens is determined by the formula:

$$\frac{1}{f} = (n-1)\left(\pm\frac{1}{R_1} \pm \frac{1}{R_2}\right), \tag{26.1}$$

where R_1 and R_2 are the radii of curvature of the front and rear spherical surfaces of the lens respectively; the plus sign is taken for a convex surface, and the minus sign is for a concave surface; n is the refractive index of the lens substance; the environment is air. The focal distance of a biconvex lens is:

$$\frac{1}{(n-1)(1/R_1+1/R_2)}.$$

The reciprocal quantity of the focal distance of a lens is called optic power D:

$$D=\frac{1}{f}. \qquad (26.2)$$

Here is the formula of a thin lens known from the school course:

$$\frac{1}{f}=\frac{1}{a_1}+\frac{1}{a_2}, \qquad (26.3)$$

where a_1 is the distance from an object to the lens; a_2 is the distance from the image to the lens.

Relation (26.1) is obtained for a thin lens with the following assumptio

1) rays that are close to the axis (***paraxial rays***) create an image;

2) rays form small angles to the principal axis of the system;

3) for all wavelengths, the refractive index is the same.

When the three conditions are met, a point image is created, i.e. each point of the object generates one point of the image.

These conditions are impossible in practice. The refractive index depen on the wavelength (dispersion).

The points of the object are aside on the optical axis, which does not meet the second condition. Use of only paraxial rays would essentially limit light flow.

All this leads to ***aberration***[1], or errors of real optical systems, reducing the quality of optical images to a great extent. However, knowing causes of aberration, it is possible to eliminate them we choose an appropriate system of lenses.

Let us consider certain basic lens aberrations.

Spherical aberration

It means that peripheral parts of a lens deflect the rays, going from point S on the axis, more than the central parts of the lens (fig. 26.1) do. The imag shape of a point of light on screen Sc is a bright spot. To eliminate spheric aberration, a system of concave and convex lenses is created.

---

[1] *Aberratio* (Latin) — deviation, deflection.

Astigmatism[1]

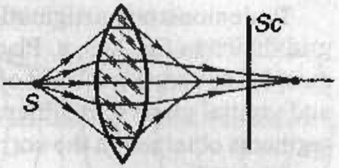


Fig. 26.1

This is a disadvantage of an optical system: a spherical light wave passing through an optical system is deformed and loses its spherical shape. There are two types of astigmatism. One of them is caused by rays falling onto the optical system and making a wide angle with the optical axis (***astigmatism of oblique beams***).

Let point *A* of the object be located beyond the principal optical axis (fig. 26.2, a); of all the rays going from this point, let us select an elementary beam so that its central ray *AS* lies in the same plane with the principal optical axis *OO'*. Any plane passing through the principal optical axis is called ***meridional***; in this case, we take the plane, in which the central ray of the elementary beam and the principal optical axis lie (In fig. 26.2, a, it is the plane of the drawing).

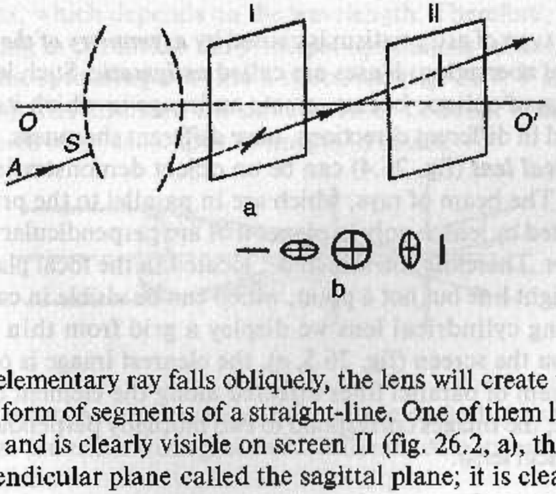


Fig. 26.2

When the elementary ray falls obliquely, the lens will create two images of point *A* in the form of segments of a straight-line. One of them lies in the meridional plane and is clearly visible on screen II (fig. 26.2, a), the other image is in the perpendicular plane called the sagittal plane; it is clearly visible on screen I. In the interval between planes I and II there is a circle of confusion which in the shape of an ellipse or circle (fig. 26.2, b). If the object is a straight-line segment, the image quality depends on its orientation. Segments located in meridional planes[2] yield a clear image in plane II, and those located in the sagittal planes do this in plane I.

---

[1] Astigmatism is not of point nature. More than one point of the image corresponds to a point of the object.

[2] The object assumed to lie in a plane, which is perpendicular to the principal optical axis.

To demonstrate astigmatism of oblique beams it is convenient to use th grid shown in fig. 26.3, a. Placing it as an object perpendicularly to the optic axis, let us create a system of segments lying in the meridional plane (rad and sagittal plane (circumferences). Fig. 26.3, b, c, shows the images of th segments obtained in the corresponding planes.

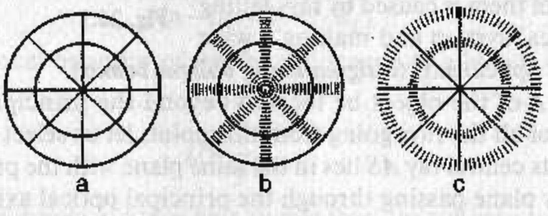


Fig. 26.3

To correct astigmatism, complex optical systems consisting of several lens are created, which makes it possible to form good images at incidence ang of 50–70°.

The other type of astigmatism is caused by ***asymmetry of the optical syste*** At this type of aberration, lenses are called ***astigmatic***. Such lenses sim[illegible]y to astigmatism of oblique beams, create an image in which its contours and lines, oriented in different directions, have different sharpness.

A ***cylindrical lens*** (fig. 26.4) can be an object demonstrating this type o astigmatism. The beam of rays, which are in parallel to the principal optic axis, is refracted by lens *L* only in planes that are perpendicular to the eleme of the cylinder. Therefore, on screen *Sc*, located in the focal plane of the lens, there is a straight line but not a point, which can be visible in case of spheric lenses. If using cylindrical lens we display a grid from thin filament with square cells on the screen (fig. 26.5, a), the clearest image is obtained in th form of a system of parallel lines directed along the element of the cylind (fig. 26.5, b, c; the images correspond to two mutually perpendicular positi of the cylindrical lens).

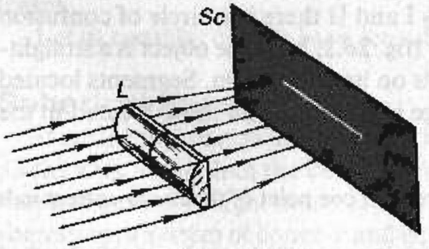


Fig. 26.4

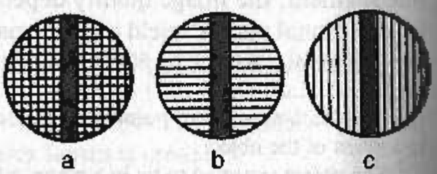


Fig. 26.5

Distortion

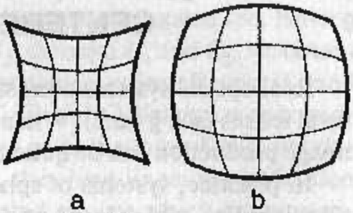


Fig. 26.6

This type of aberration occurs because of rays, sent by the object into the system, which makes wide angles to the optical axis. In this case, dependence of the linear magnification on the ray angle leads to loss of similarity between the image and the object. Typical manifestations of such a distortion are shown in fig. 26.6: case a is pincushion distortion; case b is barrel distortion; the object is a grid with square cells.

Selecting a system of several lenses with opposite nature of distortion, we can correct this aberration.

Chromatic aberration

As we can see from (26.2), the focal distance of a lens is determined by the refractive index, which depends on the wavelength. Therefore, a ray of white light propagating in parallel to the principal optical axis will be focused at its different points, spreading out into a spectrum (fig. 26.7: V are violet rays; R are red ones). The circle on the screen will be colored. This is chromatic aberration, which often accompanies images in lenses.

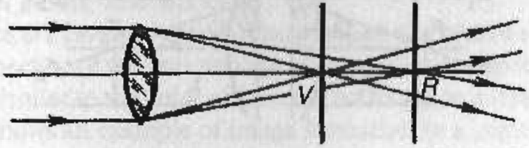


Fig. 26.7

To correct this type of aberration, achromatic optical systems are created with lenses that are made from glass with different dispersion: achromat and apochromat lenses.

Optical devices use lens systems called *anastigmats,* which correct not only chromatic aberration, but also spherical aberration as well as astigmatism.

There are other types of aberrations that we do not consider here.

Simultaneous elimination of all aberrations can present a too difficult or even unsolvable task, so they usually eliminate those errors that essentially interfere with the main purpose of the optical system. For instance, for lenses of microscopes, it is important to eliminate spherical aberration that occurs when observing objects located near the focus and represented by wide beams.

§ 26.3. CONCEPT OF A PERFECTLY CENTERED OPTICAL SYSTEM

Real optical systems consist of various components and often include several lenses (see § 26.2). When lenses are placed arbitrarily, calculation and image production can be quite complicated.

In practice, systems of spherical surfaces (lenses), whose centers lie on a straight line, which is the principal optical axis, are often applied. Such optical systems are called centered.

Let us consider certain conclusions derived from a perfectly centered optical system theory proposed by Gauss. In this system, in the space of objects each point or line is corresponded by only one point or line in the space of images. The pairs of points or lines of both spaces are called conjugate. Such systems approximate perfectly centered optical ones where paraxial rays are applied. This system is a physical abstraction that permits quantifying real systems taking into account the degree of their deviation from the ideal.

Let us point out the basic points and planes of a centered optical system (fig. 26.8; Q_1 and Q_2 are extreme spherical surfaces). These are elements of Gaussian theory; they are used to obtain images of objects.

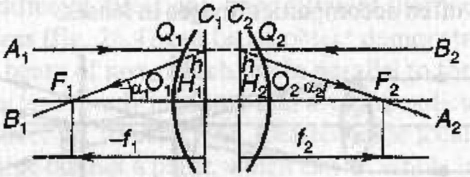


Fig. 26.8

In the space of objects, let us draw ray A_1, which is in parallel to principal optical axis O_1O_2. In the space of images, it is conjugated to ray A_2 passing through point F_2. In the space of images, point F_2 is conjugated to the infinitely remote point of the space of objects and is the second, or rear focus of the system.

Similarly, in the image space, ray B_2 is corresponded by ray B_1 of the object space passing through point F_1. In the space of objects, point F_1 conjugated to the infinitely remote point of the space of images, is the first, or front, *focus* of the system[1].

The planes passing through the foci perpendicularly to the principal optical axis are called *focal* planes.

---

[1] Since the spaces of objects and images are mutually conjugated, the concepts of "front" and "rear", "first" and "second" are conventional.

Since rays A_1 and B_1 are conjugated to A_2 and B_2 in pairs, points C_1 and C_2 of the intersections of these rays or their extensions are conjugated too. Having drawn planes, which are perpendicular to O_1O_2, through C_1 and C_2, we obtain points H_1 and H_2. We have conjugation of these planes, points H_1 and H_2, and segments C_1H_1 and C_2H_2 in pairs. Segments C_1H_1 and C_2H_2 are not only conjugated, but also equal ($|C_1H_1| = |C_2H_2| = h$) and go in the same direction relative to the optical axis (the same signs); hence, they have linear magnification of $\beta = +1$. The optical system has two conjugated planes, which are perpendicular to the optical axis, for which linear magnification of conjugate segments $\beta = +1$; such planes and the corresponding points H_1 and H_2 of the principal optical axis are called ***principal*** ones.

The distances between the foci and the corresponding principal points are called focal distances:

$$F_1H_1 = f_1, \quad F_2H_2 = f_2. \qquad (26.4)$$

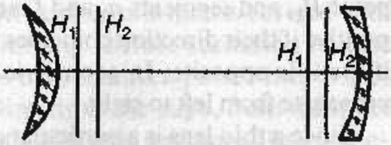


Fig. 26.9

Principal planes and points H_1 and H_2 can lie not only inside a centered optical system (fig. 26.8) but also beyond it, located absolutely asymmetrically relative to the boundary surfaces (fig. 26.9).

The principal points and foci are called ***cardinal points***, and the corresponding planes are ***cardinal planes***. These points and planes are used to form an image irrespective of the real passage of rays inside the system. The general principles are similar to the rules of forming an image in a thin lens.

Fig. 26.10 shows an example of image formation in a centered optical system. Ray *1* going from light point S_1 in parallel to the principal optical axis is extended to the intersection point with the principal plane of the image space, and then is drawn through focus F_2. Ray *2* passes through focus F_1 to the intersection point with the principal plane of the objects space, and then goes in parallel to the optical axis. The intersection point of these two rays (point S_2) forms the image of point S_1.

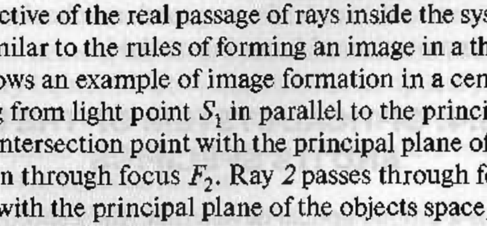


Fig. 26.10

Points N_1 and N_2 (fig. 26.10) are called nodal points. Their specific featu that conjugate rays 3 passing through them have the same inclination angle to the optical axis.

If on both sides of the optical system, there are media with equal refracti indices, the nodal points coincide with the corresponding principal points.

Thus, six cardinal points and six cardinal planes characterize any optic system.

Let us represent the most important formulas for a centered optical syste without deduction:

$$f_1/a_1 + f_2/a_2 = 1; \qquad (26.$$

$$f_1/f_2 = n_1/n_2, \qquad (26.$$

where n_1 and n_2 are refractive indices of the media on the left and right of extreme spherical surfaces of the optical system respectively; designations a_1, a_2, f_1, f_2 are explained in fig. 26.10. Segments a_1 and f_1 are counted from princip point H_1, and segments a_2 and f_2 are done from H_2. They are considered to be positive if their direction coincides with that of light propagation, and negati if the it is opposite. In geometric optics, in figures the light is assumed to propagate from left to right.

Since a thin lens is a particular case of a centered optical system, it is natural that all of the above is true for the lens too. In this case, the four cardin points — both principal and nodal ones — coincide with the center of the lens and formula (26.5) transforms into (26.3). The rules for forming an image in such lenses are known to the reader from secondary school and are also a particular case of the more universal foregoing method.

§ 26.4. OPTICAL SYSTEM OF THE EYE AND ITS SPECIFICS

A human eye is a unique optical device that fills a special place in optic This is due, first, to many optical instruments intended for visual perceptio and second, to the fact that a human (and animal) eye as a biological syste improved in the process of evolution, within the framework of bionics, and is a source of certain ideas for design and perfection of optical systems.

For physicians, the eye is not only an organ, capable of functional disorders and abnormalities, but also a source of information about certain disease not related to the eye.

Let us dwell briefly on the structure of the human eye.

The eye proper (fig. 26.11) is its eyeball with its not quite regular spherical shape: in an adult, the average anterior-posterior dimension is 24.3 mm,

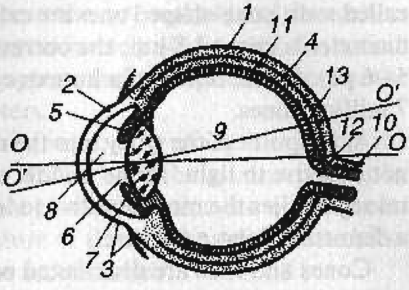


Fig. 26.11

the vertical one is 23.4 mm, and the horizontal one is 23.6 mm. The eye walls consist of three concentrically arranged membranes: external, middle, and internal one. In the anterior part of the eye, the external white membrane — sclera *1* — transits into a transparent convex cornea *2*. In the center, it is about 0.6 mm thick, in its periphery, it is up to 1 mm thick. By its optical properties, the cornea is the part of the eye with the greatest refraction index. It is like a window, through which rays of light enter the eye. The cornea radius of curvature is about 7–8 mm, the refractive index of its material is 1.38. The external layer of the cornea transits into conjunctiva *3* attached to the eyelids.

Uvea *4* is adjacent to the sclera, the inner surface of which is covered with a layer of dark pigment cells that prevent internal diffuse scattering of light inside the eye. In the anterior part of the eye, the uvea passes into iris *5*, in which there is a round hole — pupil *6*. Directly to the pupil, from the inner side of the eye, lens *7*, which is a transparent and elastic structure like a biconvex lens, is adjacent. The diameter of the lens is 8–10 mm, the radius of curvature of its anterior surface is 10 mm on average, the radius of curvature of its posterior surface is 6 mm. The refractive index of the lens material is somewhat higher — 1.41.

Between the cornea and the lens, ocular anterior chamber *8* is located, it is filled with a watery liquid, which is like water in its optical properties. Gelatinous transparent mass called vitreous body *9* occupies the entire inner part of the eye from the lens to the posterior wall. The refractive index of the vitreous body is the same as in aqueous humor.

The foregoing parts of the eye are mostly connected with its light conducting system.

Optic nerve *10* enters the eyeball through its posterior wall; ramifying, it passes into the innermost layer of the eye — retina *11*, which is the light-perceiving (receptor) system of the eye. The retina consists of several layers and varies in its thickness and sensitivity to light; it contains light-sensitive visual cells whose peripheral ends are of different shape. Oblong endings are

---

[1] The structure of the lens is like the layered structure of an onion, and the refractive index of the layers is different. Due to this specific character the lens refracts like a homogeneous material with a higher refractive index than the refractive index of any layer.

called rods, cone-shaped ones are called cones. Rods are 63–81 μm long, their diameter is about 1.8 μm, the corresponding dimensions of cones are 35 and 5–6 μm. In the retina of a human eye, there are about 130 million rods and 7 million cones.

At the point of the entry into the optic nerve, there is blind spot *12* which is not sensitive to light. In the middle of the retina, slightly closer to the temporal region, lies the most sensitive to light yellow spot *13* whose central part has a diameter of about 0.4 mm.

Cones and rods are distributed on the retina unevenly. Cones are located mainly in the central part of the retina, inside the yellow spot; in the center of the yellow spot, there are only cones, on the edges of the retina, there are only rods.

Let us first consider specific features of the light-conducting system of the eye.

The eye can be represented as a centered optical system formed by the cornea, fluid of anterior chamber and lens (four refractive surfaces), and bounded by the air medium anteriorly, and by the vitreous body posteriorly. Princip optical axis *OO* (fig. 26.11) passes through the geometric centers of the cornea, pupil, and lens. In addition, there is visual axis *O′O′* of the eye, which determines the direction of the highest sensitivity to light and passes through the centers of the lens and the yellow spot. The angle between the princip optical axis and visual axis is about 5°.

Fig. 26.12 shows focuses, principal points, planes, and nodal points of an averaged normal eye (dimensions are given in millimeters). For simplification, this system is often replaced with a *reduced* eye, i.e. a lens surrounded by the air from the side of the space of object, and liquid with $n = 1.336$ from the space of images. In a reduced eye model, the general principal plane is 1.6 mm from the anterior corneal surface; the nodal points coincide and are 7.2 mm from the corneal surface.

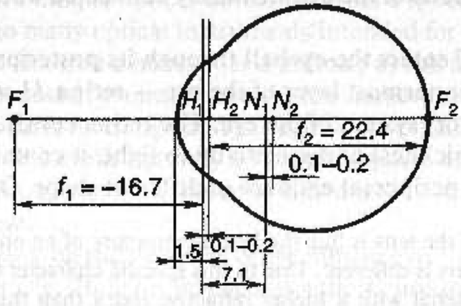


Fig. 26.12

The main refraction of light takes place at the external edge of the cornea; the optical power of the entire cornea is approximately 40 diopters; the optical power of the lens is approximately 20 diopters, and the optical power of the entire eye is approximately about 60 diopters.

At different distances, objects should give equally sharp images on the retina. From formula (26.5), we can see that this can be attained either by changing distance a_2 between the principal plane and retina the way it occurs in photo cameras or changing the curvature of the lens or, hence, focal distances f_1 and f_2. In the human eye, it is the second case.

Eye adaptation to clear vision of objects at different distances — focus adjustment — is called accommodation[1].

When an object is in infinity, in a normal eye, its image is on the retina. For all that, the lens is accommodated to infinity and its optical power is at its lowest. If an object approaches the eye, the curvature of the lens increases; the closer the object, the higher the optical power of the eye is; the optical power varies within 60–70 diopters.

When an object approaches the eye up to 25 cm, in a healthy adult accommodation occurs without stress. Due to the habit of examining things in the hands, eyes most often accommodate exactly to this distance. This is called the *distance of the best vision.*

To examine closer objects, it is necessary to strain your accommodation system. The closest location of the object from the eye, at which a clear image on the retina is still possible, is called the *near point of the eye* (the near point of clear vision). The distance to the near point of the eye increases with age; therefore, accommodation declines.

On the retina, the size of an image depends not only on the size of the object but also on its distance from the eye, i.e. on the angle at which the object is visible. In this connection, there is a concept of *viewing angle*. This is the angle between the rays going from the extreme points of the object through the coincident nodal points (fig. 26.13). The figure shows that, first, the same view

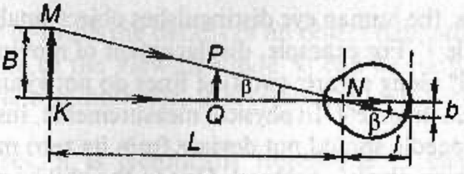


Fig. 26.13

---

[1] In certain fish, accommodation takes place due to movement of the whole lens relative to the optical axis, i.e. the first option is implemented.

angle β can correspond to different objects KM and QP, and, second, that the viewing angle defines the image size on the retina:

$$b = l\beta, \tag{26.7}$$

where l is the distance between common nodal point N and retina ($l \approx 17$ mm). Formula (26.7) is derived on the assumption that the viewing angle is small.

From fig. 26.13, it is easy to determine the connection between size B of an object, its distance L from the eye, more exactly, from the nodal points, and viewing angle β:

$$B = L\beta. \qquad (26.$$

Whence, taking into account (26.7), we obtain:

$$b = lB/L. \tag{26.9}$$

To characterize eye resolution, the ***smallest viewing angle*** is used, at which the human eye still distinguishes two points of the object. This angle is approximately equal to 1′, which corresponds [see (26.8)] to the distance of 70 μm if they are at the distance of the best vision. On the retina, in this case [see (26.7)] the size of the image is 5 μm, which corresponds to the average distance between two cones on the retina. Therefore, if on the retina an image of two points forms a line, which is shorter than 5 μm, these points will not be resolved, i.e., the eye does not distinguish them.

The same value of the smallest viewing angle will be obtained if the limitations due to diffraction of light (see § 26.8) are taken into account. Nature's expediency is striking: nothing excessive — the number of cones per the area unit of the retina corresponds to the maximum capabilities of geometric optics.

In medicine, eye resolution is estimated by ***visual acuity***. The norm is considered to be one (the standard of visual acuity when the smallest viewing angle is 1′).

When there are vision disorders, acuity is so many times less than norm as the least viewing angle is larger than one minute. If the least view angle is 4′, then visual acuity is 1 : 4 = 0.25.

In certain cases, the human eye distinguishes object smaller than those corresponding to angle 1′. For example, displacement of moving objects is visible when it reaches 20″ along an arc; two fine lines do not visually coincide at an angle of 12″ between them, etc. In physical measurements, instruments are often used, where their needle should not deviate from its zero mark (stroke) of the scale (potentiometers, electric bridges). Due to the ability of the eye to notice small line displacements, such instruments can show more accurate readings than those in which the distance between a mark and the needle is determined

---

[1] This is not only a physical problem, but also physiological one.

The eye sensitivity to light and color, as well as biophysical aspects of vision are considered in the seventh section.

§ 26.5. DEFECTS OF THE EYE OPTICAL SYSTEM AND THEIR ELIMINATION

Aberrations, which are inherent to lenses, are hardly felt by the eye.

Spherical aberration is invisible due to the smallness of the pupil becomes manifest only at twilight, when the pupil is dilated: images are blurred. Although the eye is not an achromatic system, chromatic aberration is not felt due to the selectivity of radiation visibility and small diameter of the pupil. Astigmatism of oblique beams does not take place since the eye is always set in the direction of the observed object. The only exception is astigmatism caused by asymmetry of the optical system (non-spherical shape of the cornea or lens). This manifests itself, in particular, as inability of the eye to perceive mutually perpendicular lines on the test table with equal clarity. Special glasses with cylindrical lenses compensate this visual defect.

The optical system of the eye is prone to certain specific disorders.

In the normal eye, in the absence of accommodation, the rear focus is on the retina directly; such an eye is called ***emmetropic*** or ***ametropic*** if this condition is not met.

The most common types of ametropia are ***shortsightedness*** (myopia) and ***longsightedness*** (hypermetropia). Shortsightedness is an eye disorder when it cannot provide accommodation: the rear focus is in front of the retina; in case of longsightedness, the rear focus in the absence of accommodation is behind the retina. To correct myopic vision, a divergent lens is used, a convergent lens corrects hypermetropic vision.

§ 26.6. MAGNIFYING GLASS

The possibility of resolving details of an object depends on the size of its image on the retina or the viewing angle. It can be increased moving the object closer to the eye; however, it is connected with the following limitations.

1. In some cases, it is technically impossible to change the distance between the object and the eye very much (for example, examining stars or the Sun).
2. It is impossible to bring an object closer to a distance less than the peak capabilities of the eye to the near point.

In this connection, optical devices are used to increase the viewing angl telescopes, magnifying glasses, microscopes, etc.

Let us consider the structure of a simple optical device — a magnifyi glass.

***A magnifying glass** is an optical system with the observed object placed in it anterior focal plane or in close proximity.*

An image created by the magnifying glass is in infinity or at a distan convenient for the eye. If the image is in infinity, it is visible without additional accommodation.

Fig. 26.14, a, shows with two rays the process of forming an image on th retina with a magnifying glass; *N* is the common nodal point of the optic system of the eye; the object is placed in the rear focal plane. Ray *1* pass through the center of the magnifying glass without refraction, and then is refracted by the eye. Other rays going from the same point of the object afte refraction in the magnifying glass will be in parallel to ray *1*. To determine th position of the image on the retina, we choose from these rays one that passe through the common nodal point (ray *2*). It is not refracted in the eye. Its intersection point with the retina will indicate the position of object imag To make the picture complete, it remains only to build up the initial part of ray *2* and the final part of ray *1* (shown in dotted lines).

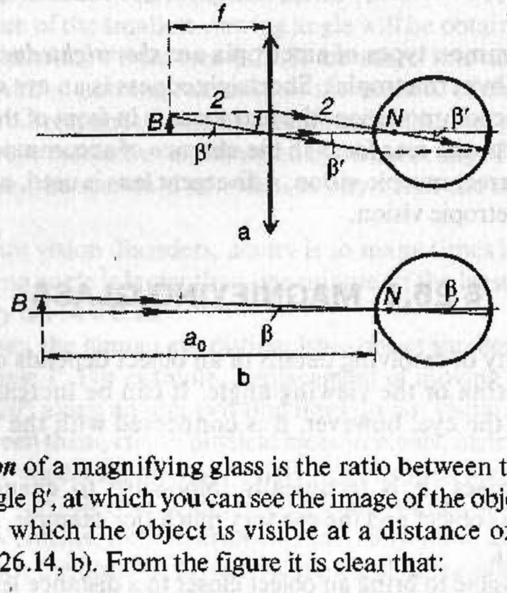


Fig. 26.14

Magnification of a magnifying glass is the ratio between the following angles: viewing angle β', at which you can see the image of the object (fig. 26.14, a), and angle β, at which the object is visible at a distance of the best visio $a_0 = 25$ cm (fig. 26.14, b). From the figure it is clear that:

$$\beta' = B/f, \quad \beta = B/a_0, \qquad (26.10)$$

where B is the linear dimension of the object. Taking into account (26.10), we obtain the magnifying glass linear magnification:

$$\Gamma = \beta'/\beta = a_0/f. \qquad (26.11)$$

We can see from it, that the fomula of linear magnification the connects constant value f of the magnifying glass with the distance of the best vision, which is a rather conventional quantity too. In the shortsighted eye, the distance of the best vision $a_s < 25$ cm, in the longsighted eye $a_l > 25$ cm, so for the shortsighted eye, the linear magnification obtained from the same magnifying glass, will be less than for the longsighted eye.

Taking into account that the stress of accommodation is very tiring for the eye and is permissible only for a short time, so when using a magnifying glass one should place the object into the focal plane, and the eye should be as close to the eyepiece as possible.

Magnifying glasses are made from one or several lenses. Magnification depends on its structure and varies in the limits from 2 to 40—50. The most common magnifying glasses have 10-X magnification.

The distance between two points resolved with a magnifying glass can be calculated by formula (26.8). For example, for 10-X magnification, we take $\beta = 0.1' \approx 3 \cdot 10^{-5}$ radians, $L = a_0 = 0.25$ m, then we obtain $B = 0.25 \cdot 3 \times \times 10^{-5} = 7.5$ μm.

§ 26.7. OPTICAL SYSTEM AND STRUCTURE OF BIOLOGICAL MICROSCOPE

To obtain high-power magnification, short-focus lenses should be used as a magnifying glass [see (26.11)]. However, such lenses are small; they have sizeable aberrations which generate limitations on magnification power. Higher magnification can be obtained examining through a magnifying glass the real image of the object created by an additional lens or lens system. Such an optical device is a *microscope.* The magnifying glass in this case is the ***eyepiece***, and an additional lens or lens system is the *object lens*.

In order for the eye not to be stressed, they try to match the image created by the objective lens with the focal plane of the eyepiece. Fig. 26.15 shows passing of the rays in a microscope, whose objective lens and eyepiece are convergent lenses, and passage of the rays in the eye.

Image A_1B_1 of object AB is created by object lens Ob. We find the former according to the rule of image formation in a thin lens. Ray *1*, which is in parallel to the principal optical axis, after refraction passes in the lens through

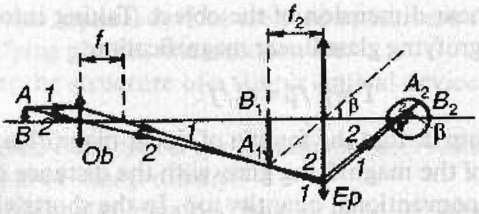


Fig. 26.15

the focus. Ray *2* passes through the lens center without refraction. A_1B_1 is located in the rear focal plane of the eyepiece.

Rays *1* and *2* reach eyepiece lenses *Ep* and are refracted by it. To show the passage of these rays after refraction in the eyepiece, let us reason in the way below.

After refraction in the lens, all rays going from a certain point of the focal plane (e.g. A_1) should propagate in parallel one to another. Let us draw ray A_1D from A_1 through the lens center; after refraction in the eyepiece, rays *1* and *2* will pass in parallel to A_1D until they reach the eye. Let ray *1* pass through common nodal point *N* of the eye and, therefore, reach point A_2 of the ret without refraction. Ray *2* will be focused in the same point. On the retina we obtain image A_2B_2 of object *AB*.

In modern optical microscopes, the object lens and eyepiece co of a lens systems, which is an ***optical system*** (fig. 26.16). The principal plan of the object lens and eyepiece of such a system are shown in the figure separately, the environment is homogeneous in respect to its refractive index. Rays *1* and *2* going from point *B* of object *AB* intersect at point *B′* where the image is formed by the object lens. Ray *2* enters the eyepiece in parallel to the principal optical axis, so it passes through focus F'_2. Since rays *1* and *2* leave the same point *B′* of the focal plane, after refraction in the eyepiece, they will be in parallel to each other.

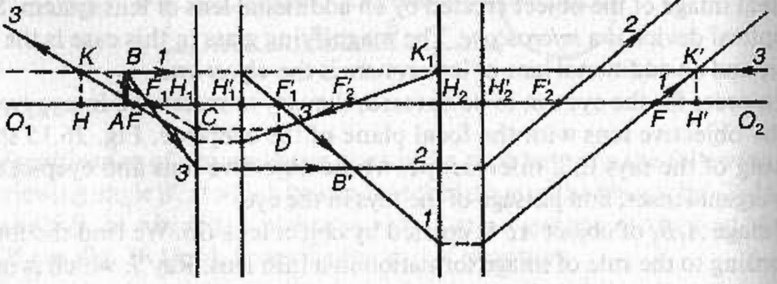


Fig. 26.16

We can specify the principal points and focuses of the microscope as a single centered optical system. Since in the object space, ray *1* is parallel to the principal optical axis, it will intersect the optical axis in rear focus F_1' in the image space.

Let us find the principal points and planes when that point and its image located in the corresponding principal planes, are equidistant from the principal optical axis.

In order not to clutter the drawing, choose point K of the anterior principal plane so that a ray going from this point in parallel to the optical axis coincides with ray *1* in the object space. Its conjugated point K' located in the rear focal plane can be found from the condition that it lies on ray *1* and is equidistant with point K from the principal optical axis. Projecting K' onto the principal optical axis, we obtain the rear principal point H'.

To find the anterior principal point from K', we direct ray *3* in parallel to the principal optical axis. It will pass through point F_2 until intersection with the rear principal plane of the object lens. To determine the direction of this ray after leaving the anterior principal plane of the lens let us make an additional construction: from point D of the focal plane, we draw ray DC in parallel to the principal optical axis; it should pass through focus F_1, and ray *3* will pass in parallel to CF_1. The intersection point of ray 3 and the principal optical axis yields anterior focus F of the microscope, and intersection with ray *1* gives us the position of point K, which is in the anterior principal plane; H is the anterior principal point of the microscope. Note that in this case, the focuses are located between the principal points.

Since the refractive indices of the medium of the object and the image spaces are equal, based on (26.6), the focal distances differ only by their sign: $f = -f'$.

Let us determine the focal distance of the microscope.

From similarity of ΔKHP and ΔCH_1F_1 and $\Delta DF_1'F_2$ and $\Delta K_1H_2F_2$ respectively, we obtain:

$$|KH|/|HF| = |CH_1|/|H_1F_1| \quad \text{or} \quad |KH|/f = |CH_1|/f_1; \tag{26.12}$$

$$|K_1H_2|/|H_2F_2| = |DF_1'|/|F_1'F_2| \quad \text{or} \quad |K_1H_2|/f_2 = |CH_1|/\Delta, \tag{26.13}$$

where f_1 is the focal distance of object lens; f_2 is the eyepiece focal distance; Δ is the distance between the rear object lens focus and the eyepiece anterior focus; Δ is called *mechanical tube length*. Dividing (26.12) by (26.13) and taking into account that $|KH| = |K_1H_2|$, we obtain $f_2 : f = \Delta : f_1$, whence the microscope focal distance is:

$$f = f_1f_2/\Delta. \tag{26.14}$$

Since the universal formula (26.11) is also valid in principle, for the micr scope [see (26.14)] we have:

$$\Gamma = \beta'/\beta = a_0/(f_1 f_2). \qquad (26.15)$$

Thus, the magnifying power of a microscope is equal to the ratio betwe the product of the mechanical tube length by the distance of the best visio to the product of the focal lengths of the object lens and eyepiece.

Formula (26.15) can be represented as a product of two factors:

$$\Gamma_{ep} = a_0/f_2 \quad \text{or} \quad \Gamma_{Ob} = \Delta/f_1, \qquad (26.16)$$

where Γ_{ep} is eyepiece magnification; Γ_{Ob} is the object lens magnification[1].

Fig. 26.17 shows a general view (a) and diagram (b) of biological microsc MVR-1. Its main parts are: base *8*, box with micro-metric mechanism *9*, obj t table *10*, objective revolver *11* with object lenses *5*, condenser *2* and eyepie *7*. The optical system consists of two parts: illumination and observation. The illuminating part includes mirror *1*, condenser *2* with a luminous-field diaphragm *3*, and removable light filter *4*, while the observation part inclu object lens, prism 6 and eyepiece connected in the microscope drawtube.

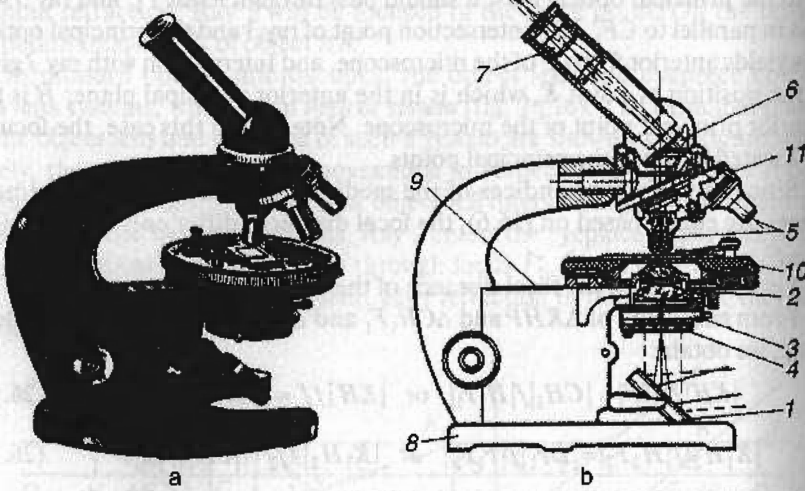


Fig. 26.17

---

[1] The reader is familiar with the formula for linear magnification of a magnifying from the course of physics at a secondary school. It should be noted that the obje located near the focus of the object lens, so the distance between the object and obj lens is approximately equal to its focal distance. Since it is small, the mechanical tub length can be considered as the distance between the object lens and the image.

From the light source, a beam of rays gets onto the mirror, which reflects it toward the diaphragm, passes through the condenser and the specimens being tested, and then gets into the object lens.

§ 26.8. RESOLUTION AND USEFUL LINEAR MAGNIFICATION OF MICROSCOPE. NOTION OF THE ABBE THEORY

From formula (26.15) it follows that after proper selection of f_1 and f_2, magnification will be arbitrarily high. However, in practice, biologists, physicians, and other specialists working with microscopes, work with magnification power greater than 1500–2000. To understand the reasons, let us familiarize ourselves with the concepts of "limit of resolution", "resolution" and "useful linear magnification of the microscope".

The limit of resolution is the smallest distance between two points of an object, when these points are distinguishable, that is visible in the microscope as two points.

Resolution is commonly known as the ability of a microscope to yield separate images of small details of the studied object being. This quantity is the reciprocal of the limit of resolution. Resolution of a microscope is determined by wave properties of light, so, for resolution limit, the expression can be obtained by taking into account the diffraction phenomena.

Let us consider the diffraction theory of microscope resolution proposed by E. Abbe.

When a transparent object is illuminated, the light scattered (diffracted) by the object gets into the microscope. A diffraction grating — an object with a quite definite structure — was chosen as the simplest object.

Let grating *D* (fig. 26.18) consist of four slits *1–4*. Secondary waves propagate from each slit; the figure shows the passage of five beams of each such wave. The secondary waves falling at equal angles to lens *L*, optical will be converged in focal plane *F*. If the path difference of secondary waves going from adjacent slits and deflected through equal angles is equal to the integer number of wavelengths, then in places marked by points on plane *F*, there will be major maxima (central, 1st, 2nd).

The pattern formed in the focal plane of the lens is called primary image. It has certain information about the object but is not an image in the conventional meaning of the word.

The image proper, or the secondary image (*1′–4′*) is formed in plane *I* due to intersection of secondary waves going from each slit. The secondary image

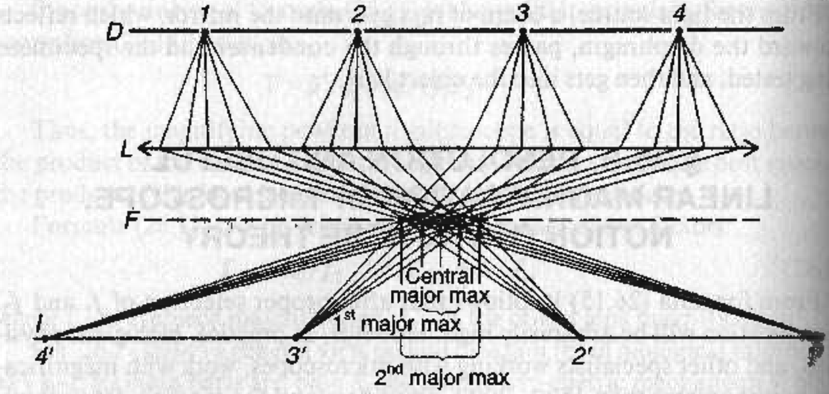


Fig. 26.18

occurs after the primary one, so it cannot contain more information abo th object than the primary image.

In optical devices, including microscopes, the length of light bea always limited, so it is important to know the following: how this will aff distortion of the object image, the minimum number of rays that can tran the correct information about the object.

The major maxima are arranged symmetrically in pairs relative to the ce tral one and, to a certain extent, duplicate each other. The set of maxim located on one side of the center together with the central one is suffici transmit information about the object. Therefore, shielding the rays go from the maxima located on the other side of the center will only reduc luminance of the object image.

When the rays going from odd principal maxima in plane *F* are shiel there occur objective conditions, at which the second major maxim fu tions as the first, the fourth one functions as the second, etc., that is [(24.29)] the image will be the same as on the diffraction grating with h of the period.

The central maximum has a common structure for gratings with diffe periods, and, therefore, it does not have information about the peculiar the object. Therefore, if we transmit only the beams of the central maxi through while screening all others, the secondary image of the obj (of the grating) will not be formed.

Abbe made such experiments with different restrictions of light bea plane *F*. He found that for the secondary image to match the object it is ne sary that the primary image transmit further at least the beams of the cent and one of the first major maxima.

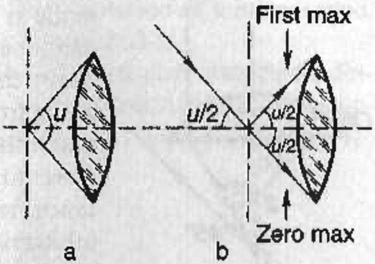


Fig. 26.19

In reality, from the object, light propagates towards the microscope object lens in a cone (fig. 26.19, a), which is characterized by its ***angular aperture***, angle u between the borderline rays of a conical light beam entering the optical system[1]. In the limiting case, according to Abbe, the borderline rays of the conical light beam will be the rays corresponding to the central (zero) and the first major maxima (fig. 26.19, b). In this case, the ray falls on the object (grating) at angle $u/2$; the angle of the first diffraction maximum is the same. From formula (24.39), $\beta = u/2$ and $\alpha = -u/2$, we obtain:

$$2c\sin(u/2) = \lambda. \tag{26.17}$$

In the above model of the object (grating), a structural parameter, which is constant c of the diffraction grating is accepted as limited resolution z, i.e. $z = c$ for α and β specified. From (26.17), we obtain:

$$z = 0.5\lambda\sin(u/2) \tag{26.18}$$

or, taking into account that $\lambda = \lambda_0/n$, and introducing $A = n\sin(u/2)$,

$$z = 0.5\lambda_0/A. \tag{26.19}$$

where A is numerical aperture; n is the refractive index of the medium between the object and the object lens.

As we can see from formula (26.19), a means of reducing the resolution limit of a microscope is to use light with a shorter wavelength. In this connection, an ultraviolet microscope is used, where micro-objects are examined in ultraviolet rays. The general optical diagram of such a microscope is like that of a conventional microscope. The main difference is, first, in using optical devices that are transparent to ultraviolet light, and, second, in specific of recording an image. Since a human eye does not perceive this radiation directly, photographic plates, luminescent screens, or electronic-optical converters are used (see section seven).

Another way to decrease the resolution limit of a microscope is to increase its numerical aperture; this is achieved increasing both the refractive index of the medium between the object and the object lens and the aperture angle. Under normal conditions (air), refractive index is equal to one. Moreover, angle $u/2$ can have high values, which can be theoretically up to 90°. If this

---

[1] It is assumed that the microscope object lens restricts the light flow most severely, i.e. it is an aperture diaphragm.

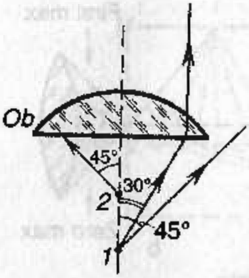


Fig. 26.20

angle is very high, the rays of the first maximu cannot enter the object lens. For example, fig. 26.2 shows that object lens *Ob* does not capture the ray going from point *1* at an angle 45°. For these rays to reach the object lens, we need to bring the obje closer to it, for example, to point *2*. However, the distance from the object to the lens cannot be chang arbitrarily, it is constant for each object lens and the object cannot be brought nearer.

The numerical aperture can be increased by means of a special liquid medium called *immersio liquid* in the interval between the lens and the cover glass of the microsco In immersion systems, in comparison with identical systems without liqui a larger aperture angle is obtained (fig. 26.21). Water ($n = 1.33$), pine nut oil ($n = 1.515$), monobromonaphthalene ($n = 1.66$) etc. are used as immersi liquid. The object lens is specially calculated for each immersion liquid, and it can only be used with this medium.

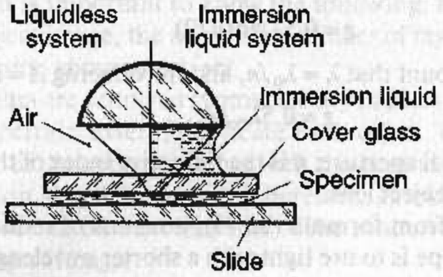


Fig. 26.21

In modern microscopes, the maximum value of angle $u/2$ is equal 70°. Such a value makes it possible to attain maximum numerical apertures and minimum resolution limits (Table 26.1).

Table 26.1

	A	*z*, μm
Liquidless system	$0.94 \cdot 1 = 0.94$	0.30
Water immersion liquid	$0.94 \cdot 1.33 = 1.25$	0.22
Oil immersion liquid	$0.94 \cdot 1.515 = 1.43$	0.19

Note. The data are given for oblique incidence of light on the object and for the wavelength of 0.555 μm.

The conditions of object lighting affect the resolution of a microscope, which is important to take into account in biological studies.

There was an amusing incident, when research biologists classified a diatom as different species, since different lighting conditions revealed the structure of its shell differently. Fig. 26.22 shows the object at full (a) and partial (b) resolution due to different lighting.

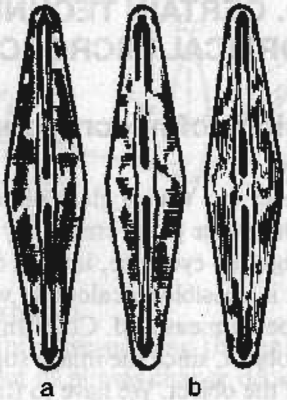


Fig. 26.22

Note that the eyepiece does not influence the resolution of the microscope at all; it only creates an enlarged image given with the object lens.

Let us estimate the *useful magnification* of the microscope using formula (26.19).

If the object has a dimension, which is equal to resolution limit z, and the length of its image is z', and if this image is located at the distance of the best vision from the eye, then magnification is:

$$\Gamma = z'/z.$$

Substituting z from (26.19) into this formula, we obtain:

$$\Gamma = Az'/(0.5\lambda_0). \qquad (26.20)$$

In the limiting case, a normal eye distinguishes two points of an object, angular distance between which is equal to 1 (see § 26.4). It is believed that convenient resolution should correspond to a viewing angle in the range from 2 to 4′ or z' (at the distance of the best viewing) from 140 to 280 μm. Substituting them together with $\lambda_0 = 0.555$ μm into formula (26.20), we obtain the interval of microscope magnification values:

$$500A < \Gamma < 1000A. \qquad (26.21)$$

These magnification values are called ***useful*** since in this case, a human ey distinguishes all the structure elements, which are resolved by the microsc

Substituting the numerical aperture of immersion oil system ($A = 1$ into (26.21), we obtain the following inequation for useful magnification for such a microscope: $700 < \Gamma < 1400$.

§ 26.9. CERTAIN TECHNIQUES OF OPTICAL MICROSCOPY

Measuring dimensions of miscroscopic objects using a microscope

An eyepiece micrometer is a round glass plate with scale marks on it. The micrometer is set in the image plane created by the objective lens.

When examined through the eyepiece, images of the object and the scal are super imposed, and it is possible to calculate what distance on the scale corresponds to the value being measured. Counting along the scale does not yet yield the length of the object, since the image superimposed on the scal is not yet equal to the size of the object. We have to find the scale interval of th eyepiece micrometer, for this purpose we use an object micrometer with a scale interval of 0.01 mm. Considering the object micrometer as an objec one brings into coincidence two scales; objec scale and eyepiece scale in the same field of view, and the scale interval of the eyepiece micrometer is further determined.

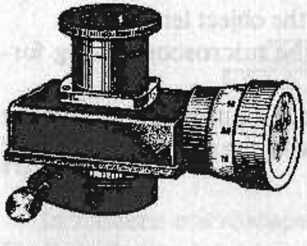


Fig. 26.23

Instead of an object micrometer, any specimen with a known length can be used, or one can resort to Goryayev's chamber. It is used for medical measurements.

Nowadays, an eyepiece screw microm the appearance of which is shown in fig. 26 is widely used. This instrument is installed instead of an eyepiece. Upon rotation, the cross-hair moves; this makes it possible to count the incremen of the micrometer.

An eyepiece screw micrometer needs preliminary calibration.

Microprojection and microphotography

Formation of a microscopic image requires the participation of a hum and results in formation of a real image in the eye.

A common microscope does not create a real image, however, a real image should be obtained for photographing (microphotography) or projecting a microscopic image onto a screen (microprojection). To do this, the image formed by object lens *Ob* should be placed farther than eyepiece *Ep* focal distance (fig. 26.24).

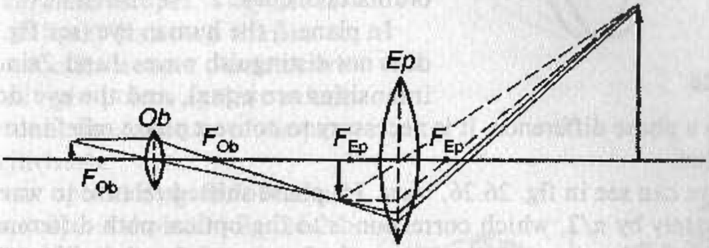


Fig. 26.24

Phase contrast method

The intensity of a light wave passing through a transparent object is approximately constant, but its phases change depending on the thickness of the object and its refractive index. In this sense, transparent objects are termed dephasing. It is impossible to see details of such objects in the conventional way. In biological studies, such objects are sometimes stained, but this can affect their properties and viability.

To make out details of dephasing objects, F. Zernike proposed the phase contrast method. Let an object consist of homogeneous transparent medium *1* with refractive index *n* where there is a transparent inclusion *2*, for example, a bacterium with refractive index n_1 (fig. 26.25). When reaching the object a part of plane-parallel beam of light will pass through the transparent object and will be focused on the small part *Φ* of the focal plane *F* by lens *L*, the other part will be diffracted by the irregularities and converged by the lens at point *A* of plane *I*.

In plane *I*, the phase composition of light oscillations is graphically shown in the intensity-phase coordinates in fig. 26.26. Curve *1* corresponds to the direct light passing through the object without diffraction; curve *2* corresponds to the light diffracted by the object. If $n_1 > n_2$, then this curve will lag in phase as shown in the figure. Curve *2* can be represented as a sum of two waves. One

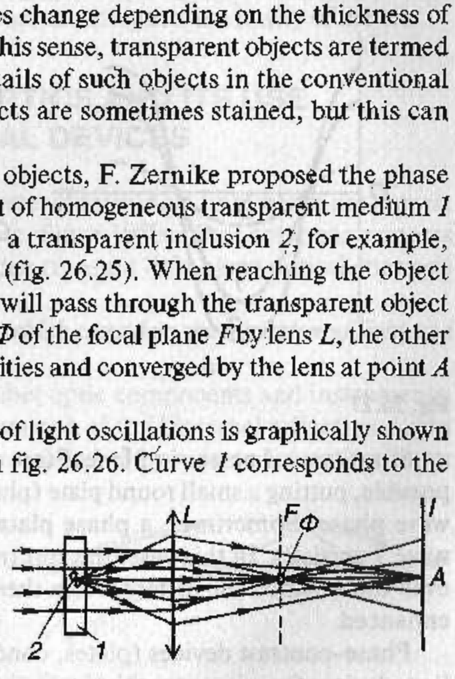


Fig. 26.25

of them (*1*) passes through the object without diffraction, the other one (*3*) is the result of diffraction by a bacterium with refractive index n_1. Curve *3* can be constructed graphically subtracting ordinates of curve *1* from ordinates of curve *2*.

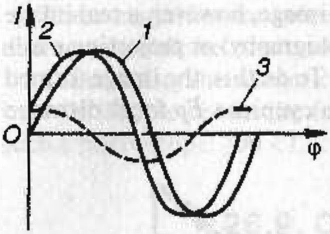
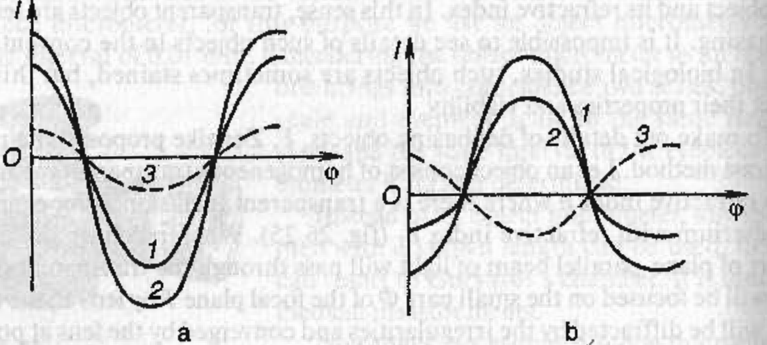


Fig. 26.26

In plane *I*, the human eye (see fig. 26.25) does not distinguish waves *1* and *2* since their intensities are equal, and the eye does not react to a phase difference. It is necessary to convert phase relief into amplitude relief.

As we can see in fig. 26.26, wave *3* is phase shifted relative to wave *1* approximately by $\pi/2$, which corresponds to the optical path difference $\lambda/4$. If we change the phase of wave *1* by $\pi/2$, then waves *1* and *3* will be either in phase (fig. 26.27, a) or in antiphase (fig. 26.27, b). We obtain curve 2 graphically as a sum of ordinates of curves *1* and *3*. The figure shows that in this case, waves *1* and *2* already differ in intensity (amplitude), therefore, the eye will see the bacterium against a homogeneous light field.



a b

Fig. 26.27

Since wave *1* passes in plane *F* (see fig. 26.25) through a small area, it is possible, putting a small round plate (phase plate) *f* in this place, to change the wave phase. Sometimes, a phase plate is made of a material that absorbs wave *1* partially. In this case, the contrast of the bacterium will be enhanced even more, since the difference in the amplitudes of waves *1* and *2* will be enhanced.

Phase-contrast devices (plates, condensers) are usually supplied as auxiliary devices for microscopes.

Ultramicroscopy

This is a method for detecting particles, whose dimensions are less than the resolution of a microscope. Microscopes working by this method are called ***ultramicroscopes***. They use lateral (oblique) lighting thanks to which, submicroscopic particles are visible as light points against a dark background; the structure of the particles is invisible.

A basic optical diagram of the ultramicroscope is shown in fig. 26.28. The light of the source falls from the left side into cell *C* together with small particles of aerosols, hydrosols, etc.; observation is carried out from the top. This method makes it possible to register particles down to 2 μm; it is used, in particular, for sanitary purposes to determine the purity of the air.

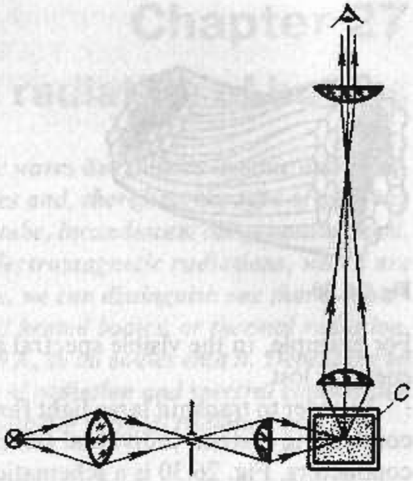


Fig. 26.28

§ 26.10. FIBER OPTICS AND ITS USE IN MEDICAL DEVICES

Traditional elements of optical systems forming a light beam are lenses, mirrors, prisms, flat-parallel plates etc. Since 1950s, fiber optic components were added to these elements, which can transmit light along optical channels called ***optical conductors***.

Fiber optics is the section of optics, which considers transmission of light and images through optical conductors.

This term is sometimes used for fiber optic components and instruments.

Fiber optics is based on the phenomenon of total internal reflection. Light getting inside a transparent fiber, surrounded by a material with a lower refractive index, repeatedly reflects and propagates along this fiber (fig. 26.29). Since in case of total reflection the reflection coefficient is relatively high (about 0.9999), the energy loss is mainly due to absorption of light by the material inside the fiber.

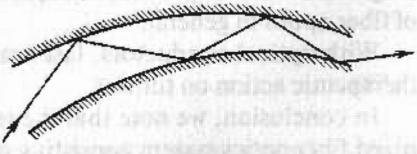


Fig. 26.29

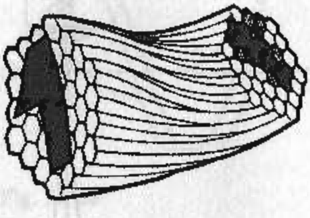


Fig. 26.30

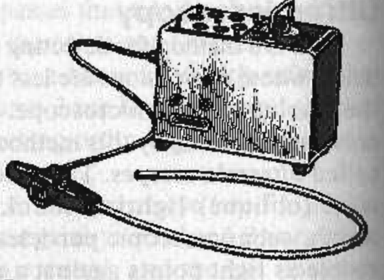


Fig. 26.31

For example, in the visible spectral range, in a fiber of 1 m long, 30–70% of energy is lost.

In order to transmit large light flows and preserve the flexibility of the ligh conducting system, individual fibers are arranged into bundles, i.e. optic conductors. Fig. 26.30 is a schematic representation of an optical conduct due to chaotic arrangement of the fibers, the image of digit 1 is distorted.

In medicine, optical conductors are used for solving the following two problems: transmission of light energy, mainly for cold lighting of internal cavitie and image transmission. In the first case, the position of separate fibers in the optical conductor does not matter, for the second case, it is essential that the location of the fibers at the input and output of the optical conductor is the same.

An example of fiber optics influence on modernization of existing medic devices is an ***endoscope***: a special instrument for examination of intern cavities (stomach, rectum, etc.). It consists of the following two princip parts: a light delivery system and examination part. Using fiber optics, it is possible, first, to transmit light from the bulb into the organ, through an optic conductor; thus, we avoid undesirable heating of this organ, which inevitab would occurr, if a light source were placed inside the cavity (this was the case with earlier endoscopes). Second, which is the most important thing, flexibility of fiber optic systems makes it possible to examine a greater part of cavity than rigid endoscopes.

Fig. 26.31 shows a ***fiber optics gastroscope***. It makes it possible not only t examine a stomach visually, but also to make the necessary photos for furth diagnostics. These particular needs of medicine motivated the developm of fiber optics in general.

With optical conductors, laser radiation is transmitted to the viscera for therapeutic action on tumors.

In conclusion, we note that the retina of the human eye is a highly organized fiber optics system consisting of about $130 \cdot 10^6$ fibers. It is probably most complex fiber optic system existing nowadays.

Chapter 27

Thermal radiation of bodies

A material radiates electromagnetic waves due to intra-atomic and intra-molecular processes. Energy sources and, therefore, the type of glow can be different: TV screen, fluorescent tube, incandescent lamp, rotting wood, firefly, etc. Of all the variety of electromagnetic radiations, which are visible or invisible by the human eye, we can distinguish one that is inherent in all bodies. This is radiation of heated bodies, or thermal radiation. It occurs at any temperature above 0 K, so all bodies emit it. Depending on the body temperature, the intensity of radiation and spectral composition change, so the eye does not always perceive thermal radiation as a glow.

§ 27.1. CHARACTERISTICS OF THERMAL RADIATION. BLACK BODY

The average power of radiation over a time longer than the period of light oscillations is taken as **radiant** ***flux*** Φ. In the SI system, it is expressed in *Watts* (W). Radiant flux emitted by 1 m^2 of a surface is called *radiant emittance* R_e. It is expressed in *Watts per square meter* (W/m^2).

A body heated emits electromagnetic waves of different wavelengths. Let us select a small wavelength range from λ to λ + dλ. The radiant emittance corresponding to this interval is proportional to the width of the range:

$$dR_e = r_\lambda d\lambda. \qquad (27.1)$$

where ***r_λ is the spectral density of radiant emittance or emissive ability of a body***, which is equal to the ratio between the radiant emittance of a narrow spectral range and the width of this spectral range (W/m^3). Radiant emittance of a body is a function of wavelength and temperature: $r_\lambda = r(\lambda, T)$.

Dependence of spectral density of radiant emittance on the wavelength is called *radiation spectrum* of the body.

Having integrated (27.1), we obtain an expression for the radiant emittance of a body:

$$R_e = \int_{\infty}^{2} r_\lambda d\lambda. \qquad (27.2)$$

(Integration limits are taken with a spare to take in account all possible thermal radiation.)

The ability of a body to absorb radiation energy is characterized by an ***absorption coefficient***, which is equal to the ratio between the radiant flux absorbed by the body and the radiant flux incident on it:

$$\alpha = \Phi_{abs}/\Phi_{inc}. \qquad (27.3)$$

Since an absorption coefficient depends on a wavelength, formula (27.3) is written for monochromatic radiant fluxes, and then this ratio determines the ***monochromatic absorption coefficient or absorption capacity of a body***:

$$\alpha_\lambda = \Phi_{abs}(\lambda)/\Phi_{inc}(\lambda). \qquad (27.3a)$$

Absorption capacity of a body is a function of the wavelength and temperature $\alpha_\lambda = \alpha(\lambda, T)$.

It follows from (27.3) that absorption coefficients can take values from 0 to 1. Black bodies absorb radiation especially well: black paper, fabrics, velvet, soot, platinum black, etc.; bodies with a white surface and mirrors absorb radiation poorly.

***A body with an absorption coefficient equal to one for all frequency ranges is called a black body.* It absorbs all incident radiation.**

There are no black bodies in nature; this concept is a physical abstraction. The model of a black body is a small hole in a closed opaque cavity (fig. 27.1).

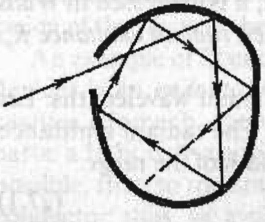


Fig. 27.1

A ray, having entered this hole, is reflected from the walls repeatedly, and will be absorbed almost completely. Further on, we will take exactly this model as the black body.

A body whose absorption coefficient is less than one and independent of the incident light wavelength is called a gray body.

There are no gray bodies in nature, but certain bodies in a certain wavelength range do their emitting and absorbing as gray bodies. For example, the human body is sometimes considered to be gray; its absorption coefficient is about 0.9 for the infrared region of the spectrum.

§ 27.2. KIRCHHOFF'S LAW

There is a definite relationship between spectral density of radiant emittance and the monochromatic coefficient of absorption, which can be explained with the following example.

In a closed adiabatic shell, there are two different bodies in conditions of thermodynamic equilibrium, and their temperatures are the same. Since the state of the bodies does not change, each of them radiates and absorbs the same amount of energy. The radiation spectrum of each body should coincide with the spectrum of electromagnetic waves absorbed by it, otherwise the thermodynamic equilibrium would be disturbed. It means that if one of the bodies emits any waves, such as red ones, to a greater extent than the other it should absorb more amount of energy too.

The quantitative connection between radiation and absorption was determined by G. Kirchhoff in 1859: *the ratio between the emissive ability of a body* **$r(\lambda, T)$ and its absorption capacity $\alpha(\lambda, T)$ is the same function $\varepsilon(\lambda, T)$ for all bodies (*Kirchhoff's law*):**

$$\frac{r(\lambda,T)}{\alpha(\lambda,T)} = \varepsilon(\lambda,T). \tag{27.4}$$

For a black body, $\alpha(\lambda, T) = 1$, therefore, $\varepsilon(\lambda, T)$ is the radiant emittance of a black body.

Using Kirchhoff's law, emissive ability of a real body can be expressed through its absorption capacity:

$$r(\lambda,T) = \alpha(\lambda,T)\cdot\varepsilon(\lambda,T). \tag{27.5}$$

Then formula (27.1) takes the following form:

$$\mathrm{d}R_e = \alpha(\lambda,T)\cdot\varepsilon(\lambda,T)\cdot\mathrm{d}\lambda. \tag{27.6}$$

For all that, the absorption capacity of the body can be determined experimentally.

§ 27.3. LAWS OF BLACK BODY RADIATION

Black body radiation has a continuous spectrum. For different temperatures, curves of radiation spectra are shown in fig. 27.2. From these experimental curves, a number of conclusions can be made.

There is maximum spectral density of radiant emittance, which is shifted along the direction of short waves when temperature increases.

Based on (27.2), black body radiant emittance R_e can be found as an area limited by the curve and the absciss axis, or

$$R_e = \int_{\infty}^{2} \varepsilon_\lambda \mathrm{d}\lambda. \tag{27.7}$$

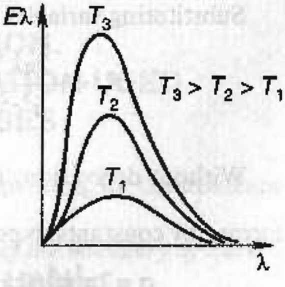


Fig. 27.2

From fig. 27.2, we can see that radiant emittance increases as the black body is heated.

For a long time, noone could establish theoretical dependence between the spectral density of radiant emittance of a black body and the wavelength and temperature, so that the dependence would match experimental data. M. Planck did it in 1900.

In classical physics, emission and absorption of radiation by a body were considered to be a continuous process.

Planck concluded that these basic concepts did not permit obtaining the correct dependence. He hypothesized that a black body radiates and absorbs energy not continuously, but in certain discrete portions — quanta. Representing a radiating body as a set of oscillators, whose energy can change only by a quantity, which is a multiple of $h\nu$, Plank obtained the following formula

$$\varepsilon_\lambda = \frac{2\pi hc^2}{\lambda^5}\frac{1}{\exp[hc/(kT\lambda)]-1} \quad \text{or} \quad \varepsilon_\nu = \frac{2\pi h\nu^3}{c^5}\frac{1}{\exp[h\nu/(kT)]-1} \tag{27.8}$$

(h is Planck's constant; c is the speed of light in vacuum; k is the Boltzmann constant), which describes the experimental curves shown in fig. 27.2 perfectly well.

Based on (27.6) and (27.8,) the spectrum of gray body radiation can be expressed by the following dependence:

$$r_\lambda = \alpha\frac{2\pi hc^2}{\lambda^5}\frac{1}{\exp[hc/(kT\lambda)]-1}, \tag{27.9}$$

where α is gray body absorption coefficient.

From formula (27.8), we can derive laws that were determined before Planck's fundamental work. Let us calculate the black body radiant emittance by formula (27.7) taking into account (27.8):

$$R_e = 2\pi hc^2\int\limits_0^\infty \frac{d\lambda}{\lambda^5\{\exp[hc/(kT\lambda)]\}-1}$$

Substituting variables $hc/(kT\lambda) = x$, we write down:

$$R_e = -\frac{2\pi k^4}{h^3c^2}T^4\int\limits_\infty^0 \frac{x^3dx}{e^x-1} = \frac{2\pi k^4}{h^3c^2}T^4\int\limits_0^\infty \frac{x^3dx}{e^x-1}. \tag{27.9a}$$

Without derivation, we write that $\int\limits_0^\infty \frac{x^3dx}{e^x-1} = \frac{\pi^4}{15}$. In (27.9a), the multiplier formed by constants, is equal to:

$$\sigma = 2\pi^5k^4/(15h^3c^2) \approx 5.6696\cdot 10^{-8}\ \mathrm{W(m^2\cdot K^4)}. \tag{27.10}$$

Finally, we obtain:

$$R_e = \sigma T^4. \tag{27.11}$$

It is the ***Stefan–Boltzmann law*: the black body radiant emittance is proportional to the fourth power of its thermodynamic temperature**. Quantity σ is called the Stefan–Boltzmann constant.

For gray bodies [see (27.2) and (27.9)], we obtain:

$$R_e = \int_0^\infty \alpha \frac{2\pi hc^2}{\lambda^5} \frac{d\lambda}{e^{hc/(kT\lambda)} - 1} = \alpha\sigma T^4. \tag{27.12}$$

Different bodies (furnace, electric stove, metal bar, etc.) can qualitatively illustrate the Stefan–Boltzmann law: as they are heated, more intensive radiation is felt.

Function ε_λ written in the form of (27.8) has an extreme value at the condition:

$$d\varepsilon_\lambda / d\lambda = 0. \tag{27.13}$$

From this, we obtain ***Wien's displacement law***:

$$\lambda_{max} = b/T, \tag{27.14}$$

where λ_{max} is the wavelength corresponding to the maximum of the spectral density of the black body radiant emittance; $b = 0.28978 \cdot 10^{-2}$ m·K is the Wien's constant. This law is valid for gray bodies also.

Wien's law manifestation can be observed in our everyday life. At the room temperature, thermal radiation of bodies is mainly in the infrared spectral region and is not perceived by the human eye. If the temperature rises, the body begins to glow in dark red light; at a very high temperature, one begins to glow in white blue hue, the feeling of body heat increases.

The laws of Stefan–Boltzmann and Wien permit taking the temperature of bodies (optical pyrometry) measuring the radiation of these bodies.

§ 27.4. SOLAR RADIATION. SOURCES OF THERMAL RADIATION USED FOR MEDICAL PURPOSES

The most powerful source of thermal radiation providing for the existence of life on Earth is the Sun.

The flux of solar radiation falling on 1 m² of area of the boundary of Earth's atmosphere is 1350 W. This quantity is called the solar constant.

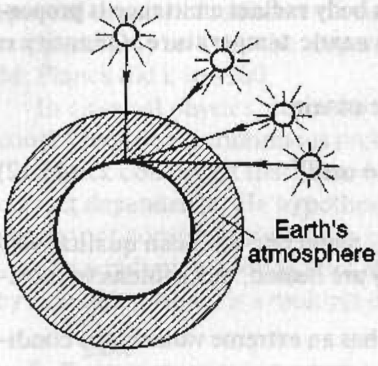


Fig. 27.3

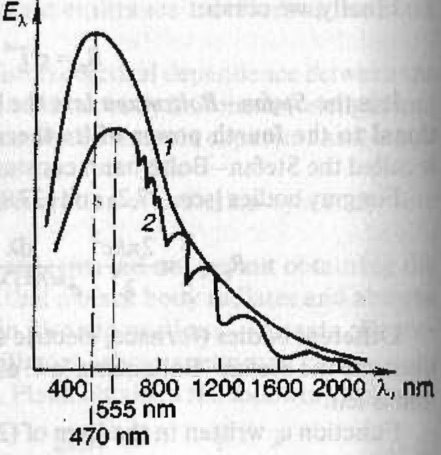


Fig. 27.4

Depending on the sun's altitude above the horizon, in the atmosphere the path of the solar rays varies in a quite broad range (fig. 27.3; the atmospheric boundary is represented schematically) with the maximum difference of 30 times. Even under the most favorable conditions, 1 m^2 of the Earth surface receives a flux of solar radiation of 1120 W. In July, in Moscow, at the highest position of the Sun, this quantity is only as much as 930 W/m^2. At other time of day, atmospheric losses are even higher.

Atmospheric extinction is accompanied by a change in spectral composition of radiation. Fig. 27.4 shows the spectrum of solar radiation at the atmospheric boundary (curve *1*) and on the Earth's surface (curve *2*) when the Sun at the highest elevation. Curve *1* is close to the spectrum of the black body; its maximum corresponds to the wavelength of 470 nm, which, according to Wien's law, makes it possible to determine the temperature of the solar surface, which is about 6100 K. Curve *2* has several absorption lines, and its maximum is about 555 nm. The intensity of direct solar radiation is measured with an ***actinometer***.

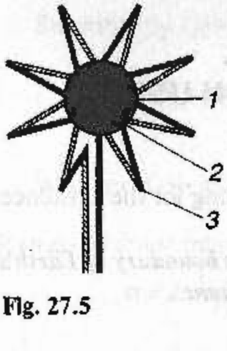


Fig. 27.5

Its operation principle is based on heating blackened surfaces of bodies by solar radiation.

In ***Savinov–Yanishevsky's*** thermoelectric actinometer (fig. 27.5) the receiving part of radiation is a thin silver disc *1* blackened from outside. Thermoelement junctions *2* are soldered to the electrically insulated disc; other junctions *3* are attached to a copper ring (not shown in the figure) inside the actinometer housing and are shaded. Under the influence of solar radiation, an electric current occurs in the thermopile (see § 15.6), the intensity of which is proportional to the radiation flux.

Dosed exposure to solar radiation is used as heliotherapy, as well as a means of developing resistance to the cold.

For therapeutic purposes artificial sources of thermal radiation are used: Incandescent bulbs (Sollux), infrared emitters (Infrarouge) fixed in a special reflector in a tripod. Infrared emitters have a structure similar to household electric heaters with a round reflector. The spiral of the heating element incandesces by a current to a temperature about 400–500 °C.

§ 27.5. DISSIPATION OF BODILY HEAT. NOTION OF THERMOGRAPHY

The human body maintains a definite temperature due to thermoregulation, an essential part of this is heat exchange between the body and the environment. Let us consider certain features of such heat exchange, assuming that the ambient temperature is lower than the temperature of the human body.

Heat exchange is implemented through heat conductance, convection, evaporation, and radiation (absorption).

It is difficult or even impossible to point out accurately the distribution of emitted heat between these processes as it depends on many factors: body state (temperature, emotional state, mobility, etc.), environmental conditions (temperature, humidity, the air motion, etc.), clothes (fabric, shape, color, thickness).

However, it is possible to make approximate and average estimates for people who do not have excessive physical activity and live in temperate climate conditions.

Since thermal conductivity of the air is low, this type of heat dissipation is negligible.

Convection is more essential, it can be not only usual, natural, but also forced when the air blows over a heated body. Clothes are of great importance in reducing convection. In temperate climates, 15–20% of human heat is dissipated through convection.

Evaporation takes place from the skin and pulmonary surfaces; it is responsible for about 30% of heat loss.

The greatest part of heat losses (about 50%) is due to radiation into the environment from exposed parts of the body and clothes. The main part of this radiation is in the infrared range with a wavelengths from 4 to 50 μm.

To calculate these losses, we make the following two basic assumptions,

1. Emitting bodies (skin, clothes fabric) are considered to be gray one This permits using formula (27.12).

Let us call the product of the absorption coefficient by the Stefa Boltzmann constant ***reduced radiation coefficient***: $\delta = \alpha\sigma$. Then (27.12) will take the form:

$$R_e = \delta T^4. \tag{27.15}$$

Below are the absorption coefficient and the reduced radiation coeffici for some bodies (Table 27.1, see below).

Table 27.1

Material	α	Δ, 10^{-8} W/(m^2 K^4)
Cotton fabric	0.73	4.2
Wool, silk	0.76	4.3
Human skin	0.90	5.1

2. Let us apply the Stefan–Boltzmann law to non-equilibrium radiatio to which, in particular, radiation of the human body belongs.

If a naked person whose body surface has a temperature T_1 is in a roo with temperature T_0, then radiation losses can be calculated in the followi way. In accordance with formula (27.15), a person emits power $P_1 = S\delta T_1^4$ from the entire active surface of his or her body of area S. At the same time th person absorbs a part of the radiation emitted by objects, room, walls, ceilin etc. If the surface of the human body had a temperature equal to the tempe ture of the air in the room, the emitted and absorbed power would be the sam and equal to $P_0 = S\delta T_0^4$.

At other temperatures of the body surface, the human body also will absor equal power.

Based on the last two equations, we obtain the power lost by the perso upon interacting with the environment through radiation:

$$P = P_1 - P_0 = S\delta(T_1^4 - T_0^4).$$

In a dressed person, T_1 is the temperature of the clothes surface. Here is a quantitative example explaining the importance of clothing.

At an ambient temperature of 18 °C (291 K) a naked person whose skin temperature is 33 °C (306 K) loses energy through radiation per second from an area of 1.5 m^2:

$$P = 1.5 \cdot 5.1 \cdot 10^{-8}(306^4 - 291^4)\ \text{J/s} \approx 122\ \text{J/s}.$$

At the same ambient temperature, in cotton clothes, the surface temperature of which is 24 °C (297 K), energy loss through radiation per second is:

$$P_{\text{cl}} = 1.5\cdot 4.2\cdot 10^{-8}(297^4 - 291^4)\ \text{J/s} \approx 37\ \text{J/s}.$$

In accordance with Wien's law, the spectral density of the radiant emittance of a human body is at its maximum when the wavelength is about 9.5 μm and the skin surface temperature is 32 °C.

Due to the strong temperature dependence on radiant emittance (the fourth degree of thermodynamic temperature), even a small increase in the surface temperature can cause such a change in the radiated power, which will be recorded by devices reliably. Let us show this quantitatively.

Let us differentiate equation (27.15): $\mathrm{d}R_e = 4\sigma T^3\cdot \mathrm{d}T$. Having divided this expression by (27.15), we obtain $\mathrm{d}R_e/R_e = 4\mathrm{d}T/T$. It means that the relative change in radiant emittance is four times higher than the relative change in the radiating surface temperature. Thus, if the surface temperature of the human body changes by 3 °C, that is approximately by 1%, radiant emittance will change by 4%.

Healthy people have a characteristic pattern of temperature distribution about different points of the body surface. Nevertheless, inflammatory processes and tumors can change the person's local temperature.

The temperature of veins depends on the actual state of blood circulation as well as on whether the limbs are cooled or heated. Thus, recording radiation from different parts of the human body surface and taking its temperature is of diagnostic importance.

Such a method called *thermography* is coming increasingly into common use in clinical practice.

Thermography is harmless, and in the future, it can become a method of preventive large-scale examinations.

Determination of the difference in body surface temperature in thermography is mainly implemented by two methods. In one case, liquid crystal sensors are used, their optical properties make them very sensitive to small changes in temperature. Placing these indicators on the patient's body it is possible to determine visually local temperature differences by their color change.

Another method is technical, based on use of thermographic cameras (see § 27.8).

§ 27.6. INFRARED RADIATION AND ITS APPLICATION IN MEDICINE

The electromagnetic radiation in the spectral range between the threshold frequency of visual light (λ = 0.76 μm) and shortwave radiation (λ = 1–2 mm) is called infrared (IR) radiation.

The infrared region of the spectrum is conventionally divided into near (0.76–2.5 μm), medium (2.5–50 μm) and far (50–2000 μm) infrared light.

Heated solids and liquids emit a continuous infrared spectrum. If we substitute the limits of ID-radiation into Wien's law, instead of λ_{max}, then we will respectively obtain a temperature of 3800–1.5 K. This means that under normal conditions, all solids and liquids are practically not only sources of IR radiation, but also have the maximum radiation in the IR-region of the spectrum. The difference between real and gray bodies does not change the main idea of the conclusion.

At a rather low temperature, radiant emittance of bodies is low. Therefore, not all bodies can be used as sources of IR radiation. In this connection, along with thermal sources of IR radiation, mercury high-pressure tubes and lasers are used; they no longer give a continuous spectrum. A powerful source of IR radiation is the Sun: about 50% of its radiation is in the IR region of the spectrum.

Methods of detection and measurement of IR radiation are mainly subdivided into two groups: thermal and photoelectric ones. An example of a thermal receiver is a thermoelement whose heating generates an electric current (see § 15.6). Photoelectric receivers are photoelectric cells, electron-optic converters, photoresistors (see § 27.8).

It is also possible to detect and record infrared radiation with photographic plates and films with a special coating.

Therapeutic use of infrared radiation is based on its thermal action. The best action is achieved with short-wave IR radiation, which is close to visible light. For therapeutic purposes special tubes are manufactured (see § 27.4).

Infrared radiation enters the body to the depth of about 20 mm, so the surface layers are heated more. The therapeutic effect is due to the occurring temperature gradient, which activates the thermoregulatory system. Enhanced blood supply of the exposed spot produces a favorable therapeutic result.

§ 27.7. ULTRAVIOLET RADIATION AND ITS APPLICATION IN MEDICINE

In the spectral region between the violet edge of visible light (λ = 400 nm) and the long-wave part of the X-ray (λ = 10 nm), the electromagnetic radiation is called ultraviolet (UV) radiation.

In the region less than 200 nm, all bodies absorb UV-radiation very much, including thin layers of the air, so it is of no interest for medicine.

The other part of the UV-spectrum is conventionally divided into three regions: *A* (400–315 nm), *B* (315–280 nm) and *C* (280–200 nm).

Hot solids emit a sizeable fraction of UV-radiation at high temperatures. Nevertheless, the maximum spectral density of radiant emittance, in accordance with Wien's law, even for the longest wavelength (0.4 μm), corresponds to 7000 K. In practice, this means that under normal conditions, the thermal radiation of gray bodies cannot serve as an effective source of powerful UV-radiation. The most powerful source of thermal UV-radiation is the Sun; at the boundary of Earth's atmosphere, 9% of the solar radiation is ultraviolet radiation.

Under laboratory conditions, electric discharge in gases and metal vapors is used as a source of UV-radiation. Such radiation is no longer thermal one and has a line spectrum.

UV-radiation is mainly measured by photoelectric receivers: photocells, photomultipliers (see § 27.8). Sensors of UV-light are luminescent materials and photographic plates.

UV radiation is necessary for ultraviolet microscopes (see § 26.8), luminescence microscopes, luminescence analysis (see § 29.7).

The main application of UV radiation in medicine is connected with its special biological effects which are due to photochemical processes (see § 29.9).

§ 27.8. PHOTOELECTRIC EFFECT AND SOME ITS APPLICATIONS

***The photoelectric effect** is a group of phenomena occurring from interaction of light with materials: it is either emission of electrons (called external photoelectric effect), change in the conductivity of a material, or occurrence of an electromotive force (internal photoelectric effect).*

The photoelectric effect demonstrates corpuscular properties of light. In this chapter, this problem is presented since a number of methods of thermal radiation detection are based on this phenomenon.

The external photoelectric effect is observed in gases in individual atoms and molecules (photoionization) and in condensed media.

The external photoelectric effect in metals is considered to be the following three processes: absorption of a photon by a conduction electron generating an increase in its kinetic energy; motion of the electron toward the surface of the body; electron emission from a metal. Einstein's equation describes this process from the energy viewpoint:

$$h\nu = A + m\upsilon^2/2, \qquad (27.16)$$

where $h\upsilon = \varepsilon$ is the photon energy; $m\upsilon^2/2$ is the kinetic energy of the electro emitted from the metal; A is the work function.

If illuminating a metal with monochromatic light we reduce the frequen of radiation (increasing the wavelength), then, starting with a certain value of the frequency, called threshold frequency, the photoelectric effect will ceas According to (27.16), the limiting case is corresponded by zero kinetic energ of the electron, which leads to the following relation:

$$h\nu_{th} = A \quad \text{or} \quad \lambda_{th} = hc/A. \tag{27.17}$$

Using these expressions, work function A is determined.

Let us represent values of the threshold frequency for the photoelect effect and the work function for some metals (Table 27.2).

Table 27.2

Metal	λ_{th}	A, eV
Silver	260	4.75
Zinc	290	4.20
Cesium	620	2.0

As we can see, the term "long-wavelength threshold of photoelectric effect" does not necessarily mean that the threshold is in the long-wavelength regio

The internal photoelectric effect is observed in case of lighting semicon ductors and dielectrics if the photon energy is sufficient to move an electro from the valence band into the conduction band. In impurity semiconducto the photoelectric effect is also present even if the photon energy is suffici to move electrons to the conduction band from donor impurity levels or from the valence band to acceptor impurity levels. In this way, photoelectric conductivity occurs in semiconductors and dielectrics.

An interesting kind of internal photoelectric conductivity can be observed upon a contact between an electronic semiconductor and hole semiconducto In this case, electrons and holes occur under the impact of light and are divided by the electric field of p–n-junction: electrons move into n-type semiconductor, and holes go into p-type semiconductor. In this case, the contact difference of potentials between the hole and electronic type semiconducto changes in comparison with the equilibrium potentials difference, that is a photo-electromotive force occurs. This form of internal photoelectric effect is called a barrier layer photoelectric effect.

It can be used for direct conversion of electromagnetic radiation energy into the energy of electric current.

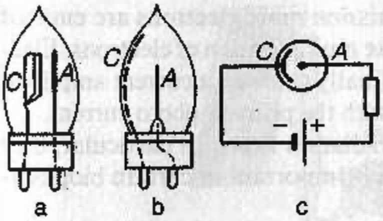


Fig. 27.6

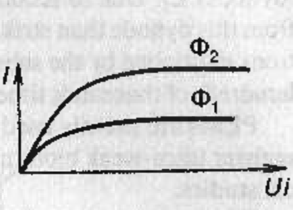


Fig. 27.7

Electro-vacuum or semiconductor devices, whose operation principle is based on the photoelectric effect, are called photoelectronic devices. Let us consider the structure of some such devices.

The most common photoelectronic device is a photocell. A photocell based on the external photoelectric effect (fig. 27.6, a) consists of an electron source — photocathode C, onto which light falls, and anode A. The whole system is enclosed in a glass bulb where there is no air. The photo-cathode, that is a photosensitive layer, can be directly applied onto a part of the internal surface of the bulb (fig. 27.6, b). In fig. 27.6, a circuit diagram of connection of the photocathode is shown.

For vacuum photocells, the operating mode is the saturation mode, which corresponds to horizontal parts of the current-voltage characteristic curves obtained at different values of the luminous flux (fig. 27.7; $\Phi_2 > \Phi_1$).

The main parameter of the photocell is its sensitivity, which is expressed by the ratio between the photocurrent intensity to the corresponding light flux. For vacuum photocells, this value is as high as 100 μA/lm.

To increase the photo-current intensity, gas-filled photocells are also used, in which a non-self-sustained dark discharge occurs in an inert gas, as well as secondary electron emission, which is emission of electrons due to bombardment of the metal surface by a beam of primary electrons. The latter is used in photo electron multipliers (PEM).

A schematic representation of PEM is shown in fig. 27.8. The photons striking cathode C emit electrons that are focused on the first electrode

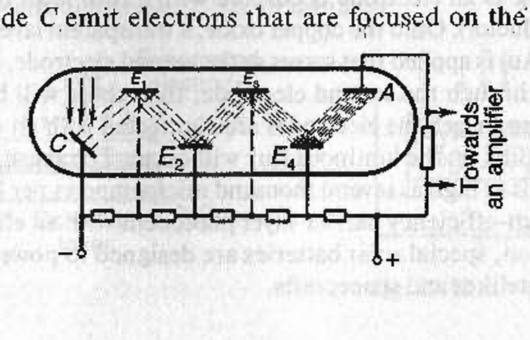


Fig. 27.8

(dynode) E_1. Due to secondary electron emission more electrons are emitt from this dynode than strike it, i.e. it looks like multiplication of electrons. Electrons multiplied by the subsequent dynodes finally generate a current amplifi hundreds of thousands times in comparison with the primary photo current.

PEMs are mainly used to measure low radiation flows, in particular, the register ultra-weak bioluminescence, which is important in certain biophy cal studies.

The external photoelectric effect is the operation principle of electro-op cal converter (EOC), which is intended to convert images from one spectr region to another, as well as to enhan the luminance of images.

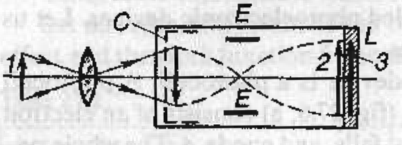


Fig. 27.9

A diagram of the simplest EOC is shown in fig. 27.9. The light image of object *1* projected onto semitranspar photocathode *C* is converted to electronic image *2*. Electrons accelerat and focused by the electric field of electrodes strike luminescent screen *L*, on which the electronic image is converted to light image *3* by cathodolum nescence again.

In medicine, EOCs are used to enhance the luminance of an X-ray imag (see § 31.4), which makes it possible to reduce the radiation exposure of patie very much. If a signal from the EOC is sent in the form of a sweep trace to a TV system, then a "thermal" image of objects can be visible on the TV screen. The body parts with different temperature differ on the screen either in colo when the image is colored, or in brightness, when the image is black and white. This device is called a thermographic camera and used in thermography (see § 27.5). Fig. 27.10 shows outward appearance of thermographic camera TV-03.

Barrier-layer photocells have an advantage over vacuum photocells because they operate without a power source.

One such photocell, which is from copper oxide, is shown in fig. 27.11. A copper plate serving as an electrode is covered with a thin layer of copper oxide Cu_2O (semiconductor). Onto the copper oxide, a transparent layer of a metal (for instance, gold, Au) is applied that serves as the second electrode. If the photoce is illuminated through the second electrode, then there will be a photo emf between them, and when the electrodes are connected with an electric circu a current depending on the luminous flux will occur. The sensitivity of barrie layer photocells is as high as several thousand microamperes per lumen.

Based on high-efficiency barrier-layer photocells with an efficiency of 15% for solar radiation, special solar batteries are designed to power the on-boa equipment of satellites and spacecrafts.

Fig. 27.10

The dependence of the photocurrent intensity on the illuminance (light flux) permits using photocells as illuminometers that are used in sanitary practice and in photography to determine actual exposure (in exposure meters).

Some barrier-layer photocells (from thallium-sulfide, germanium, etc.) are sensitive to infrared radiation, they are used to detect invisible heated bodies: they sort of expand the possibilities of vision. Other photocells (from selenium) have a spectral sensitivity, which is close to that of the human eye. This fact makes it possible to use them in automatic systems and instruments instead of the eye as objective receivers of the visible light range.

Devices, called **photo resistors**, are based on the phenomenon of photoconductivity. The simplest photoresistor (fig. 27.12) is a thin layer of semiconductor *1* with metal electrodes *2*; *3* is an insulator.

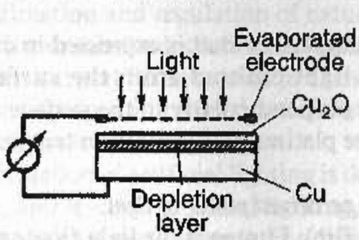


Fig. 27.11

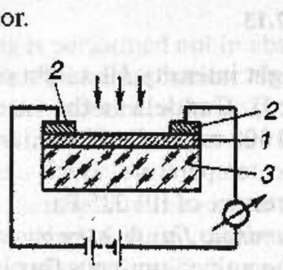


Fig. 27.12

Photoresistors, as well as photocells, make it possible to determine certain light characteristics and are used in automatic systems and measuring instruments.

§ 27.9. PHOTOMETRIC STANDARD. SOME PHOTOMETRIC QUANTITIES

Thermal radiation of bodies is widely used as a source of visible light, so let us focus on some quantities that characterize it.

To reproduce photometric quantity units with the highest attainable accuracy, the photometric standard with strictly specified geometric dimensions is used.

Its structure is sketchily shown in fig. 27.13. *1* is a tube of thorium oxide inserted into crucible *2* consisting of thorium oxide and filled with chemically pure platinum *3*; *4* is a quartz vessel with thorium oxide powder *5*; *6* is inspection window; *7* is a photometric unit that makes it possible to equalize the illuminance generated on plate *9* with standard emitter and its copy; *8* is a special electric incandescent bulb (reference standard).

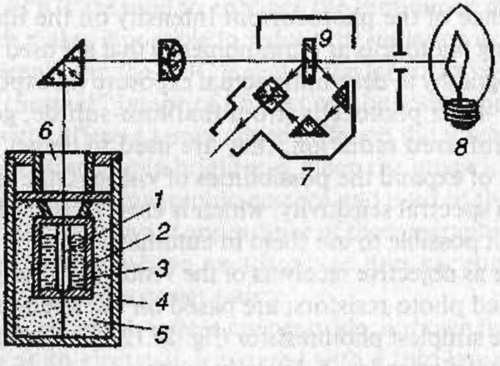


Fig. 27.13

Light intensity *I* is a light source characteristic that is expressed in candelas (cd). Candela is the intensity of light emitted from the surface of 1/600 000 m^2 of the Planckian radiator perpendicularly to the surface at the emitter temperature, which is equal to the platinum solidification temperature at a pressure of 101 325 Pa.

Luminous flux *Φ is the measure of the perceived power of light.*

The unit of luminous flux is a lumen (lm). Lumen is the light flow emitted by a point source across a solid angle of 1 steradian at light intensity of 1 cd.

***Luminous emittance** is the quantity equal to the ratio between luminous flux emitted by a radiant surface and the area of this surface:*

$$R = \Phi_{\text{rad}}/S. \tag{27.18}$$

The unit of luminous emittance is a lux (lx), which is lighting of a surface area of 1 m^2 with luminous flux of incident radiation equal to 1 lm.

To estimate the emission or reflection of light in a given direction a light characteristic called ***luminance*** is introduced. Luminance is defined as the ratio between light intensity dI of elementary surface element dS in the given direction to the projection of the emitting surface onto the plane, which is perpendicular to this direction:

$$L = \frac{dI}{dS_0} = \frac{dI}{dS\cos\alpha}, \tag{27.19}$$

where α is the angle between the perpendicular to the emitting surface and the given direction (fig. 27.14).

The unit of luminance is *candela per square meter* (cd/m^2). The photometric standard under conditions formulated above corresponds to the luminance of $6 \cdot 10^5$ cd/m^2.

Sources with equal luminance in all directions are called *lambertian* source; strictly speaking, such a source could be only the black body.

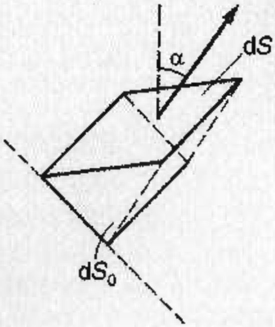


Fig. 27.14

***Illuminance** is the quantity equal to the ratio between the luminous flux incident on the given surface and the area of the surface:*

$$E = \Phi_{\text{inc}}/S. \tag{27.20}$$

In hygiene, illuminance is used to estimate lighting. Illuminance is measured with an illumination photometer. Its principle of operation is based on the photoelectric effect (see § 27.8).

Estimation and regulation of natural lighting is performed not in absolute units, but in relative indices of the coefficient of natural illumination: the ratio between natural illumination at the studied point indoors to the coincident value of outdoor illumination on a horizontal surface in the open air in the absence of direct sunlight.

Estimation of artificial lighting is done by measuring illuminance and luminance, and levels of artificial lighting are regulated by taking into account the nature of visual performance. For different operations, the limits of permissible illuminance are from one hundred to several thousand lux.

Section 7

PHYSICS OF ATOMS AND MOLECULES. INTRODUCTION TO QUANTUM BIOPHYSICS

Until the end of the 19th century, an atom was considered as an indivisible particle. However, discovery of electrons themselves and their presence in all materials convinced scientists of a complex structure of atoms. Rutherford's famous experiments on scattering of alpha particles became of decisive importance in understanding the structure of an atom. Atomic physics now had all prerequisites for its development: this branch of physics studies the structure and state of atoms, i.e. the theory of atoms, atomic (optical) spectroscopy, X-ray spectroscopy, radio spectroscopy and other problems. The physics of atoms and especially physics of molecules have a lot in common with problems considered by chemistry. There are no clear boundaries in the spheres of application of these sciences. A physician should have an idea of the nature of physical and physico-chemical processes taking place in the human body. Ultimately, these processes develop at the molecular level, so, in this, we consider problems related to energy transformation of molecules in biological systems (chemilluminescence, photobiological phenomena, etc.). These subjects are covered by the term "quantum biophysics", apparently in connotation with quantum mechanics.

Chapter 28

Wave properties of particles. Introduction to quantum mechanics

Quantum mechanics is a modern theory that establishes a way of describing and laws of motion of micro-particles (elementary particles, atoms, molecules, and nuclei) and their systems. In comparison with the classical physics, the quaint nature of quantum mechanical concepts paved the way for rejecting basic physical models, which used to look obvious and inviolable. This mainly affected the concept of a particle and the principles of its motion.

This chapter gives a notion not only of quantum mechanics, but also of the ideas and experiments that led to the origin of this theory. Electron microscopy is considered to be a method based on wave properties of electrons.

§ 28.1. DE BROGLIE'S HYPOTHESIS. EXPERIMENTS WITH DIFFRACTION OF ELECTRONS AND OTHER PARTICLES

In development of quantum mechanics, an important stage was discovery of the wave properties of micro-particles. The idea of wave properties was originally hypothesized by a French physicist Louis de Broglie (1924)[1].

The theory that light is an electromagnetic wave dominated in physics for many years. However, after the works of Planck (thermal radiation), Einstein (photoelectric effect), and other scientist, it became evident that light has corpuscular properties too.

To explain certain physical phenomena, it is necessary to consider light as a flow of particles, i.e. photons. Corpuscular properties of light do not reject but complement its wave properties. Therefore, a photon is an elementary particle of light with wave properties.

---

[1] De Broglie's hypothesis was formulated before there were any experiments confirming wave properties of particles. De Broglie writing about it later, in 1936, said the following: "...can we not assume that the electron is as dual as light? At first sight, such an idea seemed quite impertinent. You know that we always understood that the electron was an electrically charged material point, which meets the laws of classical dynamics. An electron never manifested its wave properties directly, such as, for instance, light does in its interference and diffraction phenomena. An attempt to attribute wave properties to an electron, when there was no experimental evidence for it, could look like a non-scientific fantasy".

It is logical to assume that other particles — electrons, neutrons — have wave properties too.

The formula for the photon momentum:

$$p = h\nu/c = h/\lambda \tag{28.1}$$

was used for other micro-particles with mass m moving at speed υ:

$$p = m\upsilon = h/\lambda,$$

from which

$$\lambda = h/(m\upsilon). \tag{28.2}$$

According to de Broglie, particle motion, like that of electrons, is like a wave process with length λ determined by formula (28.2). These waves are called *de Broglie waves.*

De Broglie's hypothesis was so unusual that many prominent contemporary physicists did not attach any importance to it. Several years later this hypothesis was proved experimentally: electron diffraction was discovered.

Let the electron receive kinetic energy passing through difference of potentials U, then the following equation is valid:

$$m\upsilon^2/2 = eU.$$

Let us express the velocity and substituting it into (28.2), we obtain:

$$\lambda = h/\sqrt{2emU}. \tag{28.3}$$

To obtain a quite good electron beam for recording, for example, on an oscilloscope screen, accelerating voltage of about kV is required. In this case, from (28.3), we obtain $\lambda = 0.4 \cdot 10^{-10}$ m, which corresponds to the wavelength of X-ray radiation.

In Chapter 24, it was noted that X-ray diffraction is present in crystalline bodies; therefore, to obtain electron diffraction, it is also necessary to use crystalline materials.

For the first time, diffraction of electrons by a nickel single crystal was observed by C. Davisson and L. Germer observed the J.P. Thomson and P.S. Tartakovsky (independently) did it with metal foil (polycrystalline body).

Fig. 28.1 shows an electron diffraction pattern obtained due to interaction between electrons and polycrystalline foil.

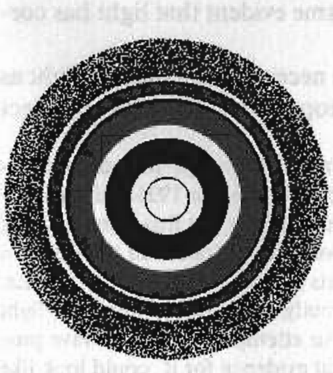


Fig. 28.1

Comparing this figure with fig. 24.21 we can see common features between electron diffraction and X-ray diffraction.

Other particles also have the ability to diffract, for instance, charged particles (protons, ions, etc.) and neutral particles like neutrons, atoms, molecules.

Particle diffraction can be used for determining ordered or disordered arrangement of atoms and molecules of materials and to estimate parameters of crystal gratings the way it is done in X-ray crystallography.

Nowadays, methods of electron and neutron diffraction are widely used.

The following questions can arise: what happens to individual particles? How are the maxima and minima formed upon diffraction of separate particles?

Diffraction experiments with very low energy electron beams, that is, somewhat like separate particles, have shown that an electron is not "spread" in different directions but behaves as an integral particle. However, the probability of electron deflection in individual directions because of interaction with an object causing diffraction varies. The chance of electrons hitting those places is most probable, if these places are calculated to correspond to the maxima of diffraction; getting into the minima places is less probable. Thus, wave properties are inherent not only to a beam of electrons, but also to each electron taken separately.

§ 28.2. ELECTRON MICROSCOPE. NOTION OF ELECTRON OPTICS

Wave properties of particles can be used not only for diffraction crystallography, but also to obtain enlarged images of objects.

It follows from (26.19) that the resolution limit of an optical microscope is mainly determined by the limiting quantity of light wavelength percived by the human eye. Substituting the expression for the wavelength of the de Broglie wave (28.3) in this formula, we will find the limit of resolution of an electron microscope, in which electron beams form the image of an object:

$$z = 0.5h/[\sqrt{2emU} \cdot \sin(u/2)]. \qquad (28.4)$$

As can be readily appreciated, the limit z of resolution of an electron microscope depends on accelerating voltage, and one can reduce the voltage considerably, while increasing the resolution exceeding that of an optical microscope.

An electron microscope and its individual components have the same function as an optical microscope, so let us use an analogy to explain its structure and principle of operation. Diagrams of both microscopes are shown in fig. 28.2 (a optical microscope, electron microscope).

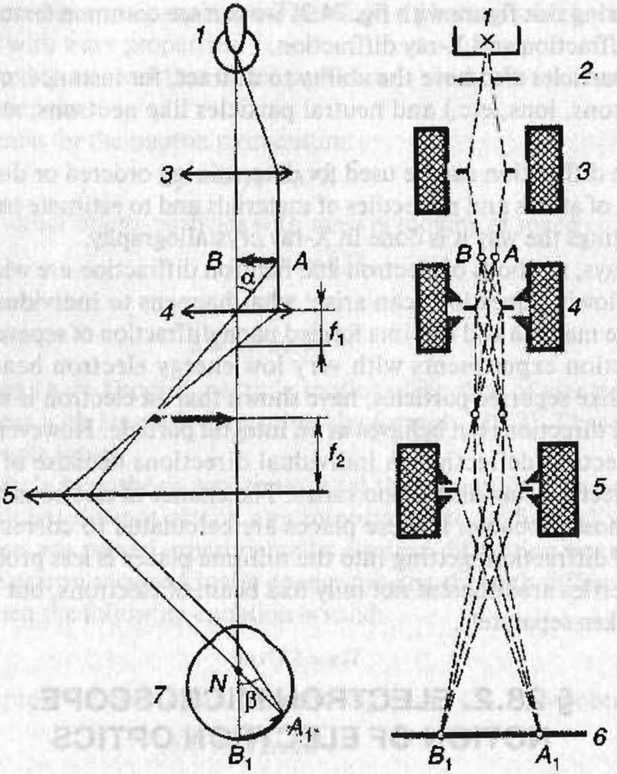


Fig. 28.2

In an optical microscope, the carrier of information of object AB is a photon as a particle of light. A light source is usually incandescent bulb *1*. Aft interaction with an object (absorption, dispersion, diffraction), the phot flow is transformed and contains information about the object. Optical devic (mainly lenses) generate a photon flow: condenser *3*, objective lens *4*, eyepiece *5*. Image A_1B_1 is visible by eye *7* (or is recorded onto a photo plate, photoluminescent screen, etc.).

In an electron microscope, an electron is a carrier of information about th object, and heated cathode *1* is a source of electrons. Acceleration of electro and formation of a beam is made using a focusing electrode and an anode; th system is called electron gun *2*. After interacting with the object (mainly dispersion), the electron flow is transformed and contains information about th object. The electron flow formation occurs under the impact of of an electr field (a system of electrodes and capacitors) and magnetic field (a syste of inductors with a current). These systems are called electronic lenses by analogy with optical lenses that form a light flux (*3* is condenser lens; *4* is electron lens serving as an objective lens; *5* is the projection lens). The image is recorded onto an electron sensitive photographic plate or cathode thermoluminescent screen *6*.

To estimate the resolution limit of an electron microscope, we substitute the accelerating voltage of 100 kV in formula (28.4) and the angular aperture about 10^{-2} rad (approximately such angles are used in electron microscopy). Then we obtain $z \approx 0.1$ nm, which is hundreds times better than in optical microscopes. Use of accelerating voltage, which is higher than 100 kV, although increasing the resolution, is associated with certain difficulties, in particular destruction of the object under study by electrons with high velocity. In practice, even with the best electron microscope, it is possible to attain a resolution limit of about 10^{-10} m; this is hundreds times better than in optical microscopes.

Advantages of an electron microscope are high resolution, making it possible to examine large molecules, the ability to change the accelerating voltage and, therefore, the resolution limit, and a relatively convenient control of the electron flow by magnetic and electric fields.

Let us specify certain features of operation of an electronic microscope. In those components, where electrons move, we have to have vacuum, otherwise, collisions of electrons with molecules of the air (gas) will lead to distortion of the image. This requirement for electron microscopy complicates the study process, makes the equipment more bulky and expensive. Vacuum distorts the native properties of biological objects; sometimes it can even destroy or deform them.

Only very thin sections are examinable with an electron microscope, because materials absorb and disperse electrons very much. Therefore, in certain cases, it is expedient to make an impression of the object surface under study on a thin layer of plastic. This action is called replication, and the plastic copy of the surface is called a *replica*.

Modern Russian electron microscope EVM-100 LM (fig. 28.3) yields a maximum 600 000-fold linear magnification and ensures a resolution limit of $3 \cdot 10^{-10}$ m. Fig. 28.4 shows images of RNA molecules in different actual states obtained with electron microscope with 100 000x magnification.

In both photons, electrons and other particles, the presence of wave and corpuscular properties simultaneously makes it possible to extend a number of concepts and laws of optics to description of motion of particles charged in electric and magnetic fields.

This analogy has made it possible to distinguish *electron optics* as an independent section of physics studying the structure of charged particle beams

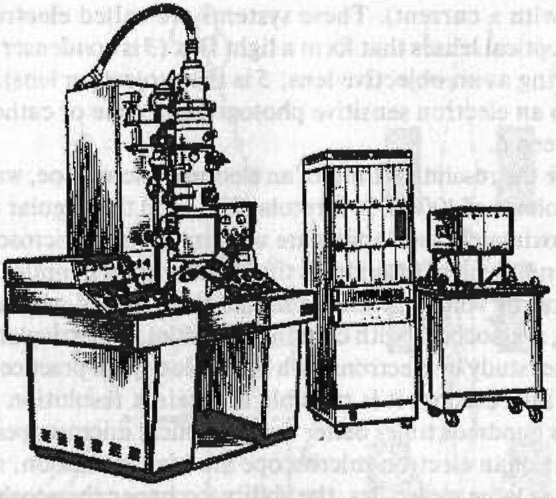


Fig. 28.3

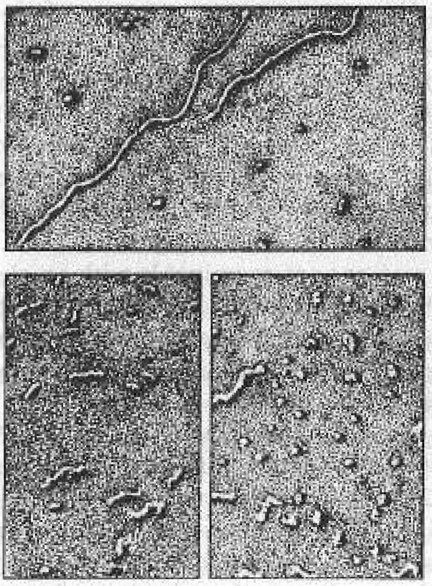


Fig. 28.4

interacting with electric and magnetic field. Similarly to conventional optic electron optics can be subdivided into *geometric* (ray) optics and *wave* (physical) optics.

Geometric electron optics describes, in particular, motion of charged particles in electric and magnetic fields. A schematic representation of image formation in an electron microscope (see fig. 28.2, b) is based on geometric electron optics.

The approach of wave electron optics is valid when charged particles show wave properties. A good illustration is determination of the resolution (resolution limit) given at the beginning of this section.

§ 28.3. WAVE FUNCTION AND ITS PHYSICAL MEANING

Since a micro-particle is correlated with a wave process that corresponds to its motion, in quantum mechanics, the state of particles is described by a wave function that depends on coordinates and time: $\psi(x, y, z, t)$.

If the field of force acting on a particle is stationary, i.e. time-independent, then ψ-function can be represented as a product of two factors, one of which depends on time, and the other one does on coordinates:

$$\psi(x, y, z, t) = f(t)\psi(x, y, z). \tag{28.5}$$

Further on we will consider only stationary states; ψ-function is a probabilistic characteristic of a particle's actual state. Let us explain the meaning of this statement.

Let us select a quite small volume $dV = dx\,dy\,dz$ in space, inside which values of ψ-function can be considered to be equal. The probability that the particle is in this volume dW_B is proportional to the volume and depends on the square of the ψ-function absolute value:

$$dW_B = |\psi|^2\,dV. \tag{28.6}$$

Hence the physical meaning of the wave function:

$$|\psi|^2 = dW_B/dV. \tag{28.7}$$

The square of the absolute value of the wave function is equal to its probability density, i.e. the ratio between the probability of finding a particle in the volume to this volume itself.

Integrating expression (28.6) by certain volume V, we find the probability of finding a particle in this volume:

$$W_B = \int_V |\psi|^2\,dV. \tag{28.8}$$

§ 28.4. UNCERTAINTY RELATIONS

An important concept of quantum mechanics is the uncertainty prin formulated by W. Heisenberg.

Let the position and momentum of a particle be measured simultaneou in this case, in determination of the abscissa and the projection of the momentum onto the abscissa axis, the errors are equal Δx and Δp_x respectively.

In classical physics, there are no restrictions prohibiting measuring both one and the other value simultaneously with any degree of accuracy, i.e. $\Delta x \to 0$ and $\Delta p_x \to 0$.

In quantum mechanics, the situation is fundamentally different: Δx and Δp_x corresponding to simultaneous determination of x and p_x, are connected by a dependence:

$$\Delta x \Delta p_x \geq h/(2\pi). \qquad (28.9)$$

Thus, the more accurately coordinate x ($\Delta x \to 0$) is determined, the les accurately projection p_x ($\Delta p_x \to \infty$) is defined, and vice versa. Similarly,

$$\Delta y \Delta p_y \geq h/(2\pi); \quad \Delta z \Delta p_z \geq h/(2\pi). \qquad (28.10)$$

Formulas (28.9), (28.10) are called uncertainty relations. Let us expla them with one model experiment.

In § 24.5, it was noted that during diffraction, a decrease in the width of a slit leads to widening the central maximum. The same phenomenon will be seen in electron diffraction by a slit in the model experiment[1]. Narrowing a slit means a decrease in Δx (fig. 28.5); this leads to a greater "spread" of the electron beam that is, to a higher uncertainty in determination of momentum and velocity of the constituent particles.

Another uncertainty relation can be represented in the following form:

$$\Delta E \Delta t \geq h/(2\pi), \qquad (28.11)$$

where ΔE is energy uncertainty of a certai state of a system; Δt is the time interval during which it exists. Relation (28.11) means that the shorter the lifetime of a system state, the more uncertain its energy quantity is. Energy levels E_1, E_2 etc., have a certain width (fig. 28.6) depending on the time of the

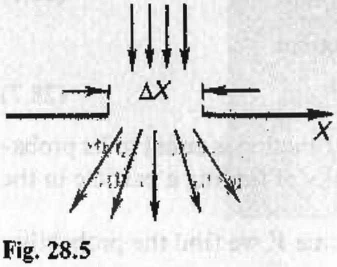


Fig. 28.5

---

[1] It is impossible to implement such an experiment in reality, since the dimensio of the slit should be about the dimensions of an atom; therefore, we describe a certai virtual experiment.

system remaining in the state corresponding to this level.

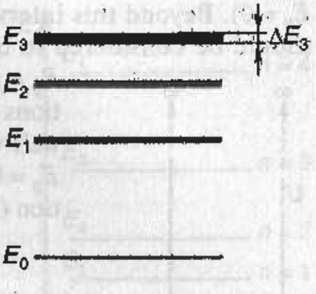


Fig. 28.6

"Blurring" of levels leads to energy uncertainty of ΔE of the emitted photon and its frequency $\Delta\nu$ when the system transits from one energy level to another:

$$\Delta E = \Delta(h\nu) = h\Delta\nu. \tag{28.12}$$

This manifest as broadening of spectral lines.

§ 28.5. SCHRÖDINGER EQUATION. ELECTRON IN A POTENTIAL WELL

Since a state of a micro-particle is described by ψ-function, then it is necessary to specify the method of finding this function taking into account external conditions. This was made possible by solving the basic equation of quantum mechanics proposed by E. Schrödinger (1926). In quantum mechanics, this equation is postulated the way Newton's second law is done in classical mechanics.

As applied to stationary states, the Schrödinger equation can be written as:

$$\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+\frac{\partial^2\psi}{\partial z^2}+\frac{8\pi^2 m}{h^2}(E-E_p)\psi=0, \tag{28.13}$$

where m is particle mass; E and E_p are its total and potential energy (the potential energy is determined by the field of force, in which the particle is located, and, for the stationary case, does not depend on time).

If the particle moves only along a certain line, for example, along the X-axis (one-dimensional case), then the Schrödinger equation is simplified very much and takes the form:

$$\frac{d^2\psi}{dx^2}+\frac{8\pi^2 m}{h^2}(E-E_p)\psi=0. \tag{28.14}$$

A most common example of using the Schrödinger equation is solving the problem of particle motion in a one-dimensional potential well.

Let an electron move along axis X only within limits $0 < x < l$ (fig. 28.7). It means that in the specified interval, ψ-function is other than zero, and beyond interval ($x \le 0$, $x \ge l$) it is equal to zero.

Since force fields do not act on the particle in the chosen interval, its potential energy can have any constant value (it is most convenient to assume

$E_p = 0$). Beyond this interval, there is no electron, so its potential energ should be considered to be infinite. Fig. 28.7 shows dependence curv

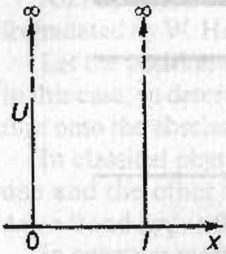


Fig. 28.7

$E_p = f(x)$. Interval $0< x < l$ meeting the above conditions is called one-dimensional rectangular r potent well with infinitely high walls. Taking into accou $E_p = 0$, for the interval $0< x < l$, the Schrödinger equation (28.14) takes the form:

$$\frac{d^2\psi}{dx^2} + \frac{8\pi^2 m}{h^2} E\psi = 0. \qquad (28.14a)$$

Introducing designation

$$\omega^2 = 8\pi^2 mE/h^2, \qquad (28.15)$$

we obtain:

$$\frac{d^2\psi}{dx^2} + \omega^2\psi = 0. \qquad (28.16)$$

This equation is analogous to the differential equation of harmonic oscillation (see § 7.1), the solution of which is:

$$\psi = \psi_0 \cos(\omega x + \varphi_0), \qquad (28.17)$$

where ψ_0 is the wave function amplitude; φ_0 is its initial phase.

To find two constants ψ_0 and φ_0, as well as possible values of ω or E, let u consider the following boundary conditions.

1. When $x = 0$, $\psi_0 = 0$. Substituting these values into (28.17), we obtain $0 = \psi_0 \cdot \cos(0 + \varphi_0) = \psi_0 \cdot \cos\varphi_0$. In this, only one value has a physic meaning: $\cos\varphi_0 = 0$, whence, $\varphi_0 = \pi/2$;
2. When $x = 1$, $\psi = 0$. Taking into account that from (28.17) $\varphi_0 = \pi/2$, we have $0 = \psi_0 \cos(\omega l + \pi/2)$. In this, only one value has a physical meanin $\cos(\omega l + \pi/2) = 0$, or $\omega l + \pi/2 = (2n + 1)(\pi/2)$, whence:

$$\omega l = n\pi/l, \qquad (28.18)$$

where n is an integer and takes values 1, 2, 3,..; $n \neq 0$, as otherwise $\psi = 0$ for any x, it would mean no electron in the potential well. Number n is called the *principal quantum number*. From (28.15), we find energy $E = h^2\omega^2/(8\pi^2 m)$, which taking into account (28.18) yields:

$$E_n = [h^2/(8ml^2)] \cdot n^2. \qquad (28.19)$$

Subscript n at E shows that to different values of principal quantum number n, a different energy corresponds too.

Substituting ω from (28.18) into (28.17) and taking into account that $\varphi_0 = \pi/2$, we obtain:

$$\begin{aligned}\psi &= \psi_0 \cos(nx\pi/l + \pi/2) = \\ &= \psi_0 \cos\pi(xn/l + 1/2).\end{aligned} \qquad (28.20)$$

Let us analyze expressions (28.19) and (28.20).

First, it is remarkable that for an electron in a potential well without any additional postulates, the solution of the Schrödinger equation leads to discrete, quantized values of energy:

$$E_1 = h^2/(8ml^2), \;\; E_2 = [h^2/(8ml^2)]\cdot 4, \text{ etc.}$$

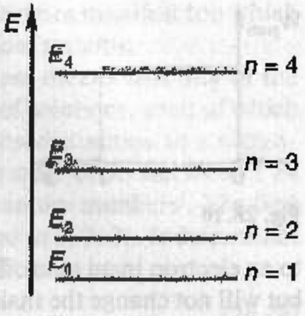


Fig. 28.8

Energy levels E_1, E_2, E_3, E_4 corresponding to different states of an electron are shown sketchily in fig. 28.8. Let us calculate the energy difference of adjacent levels $n + 1$ and n:

$$\begin{aligned}\Delta E = E_{n+1} - E_n &= h^2(n+1)^2/(8ml^2) - h^2 n^2/(8ml^2) = \\ &= h^2(n^2 + 2n + 1 - n^2)/(8ml^2) = h^2(2n+1)/(8ml^2).\end{aligned} \qquad (28.21)$$

From (28.21), we can see that at a fixed quantity of n, discreteness, i.e. difference in the energies of adjacent levels, is the smaller, the higher the size of the potential well. For example, let us calculate the following two cases for $n = 1$.

1. $l = 5\cdot 10^{-10}$ m, which approximately corresponds to the dimensions of an atom; then $\Delta E = 4.5$ eV. By the order of the magnitude, this coincides with the values obtained for the hydrogen atom according to Bohr's theory.
2. $l = 10^{-1}$ m: this actually corresponds to such a width of the potential well that the electron can be considered to be free; in this case, $\Delta E = 1.1\cdot 10^{-16}$ eV. In this, discreteness is negligible, and, essentially it can be considered that the electron energy changes continuously.

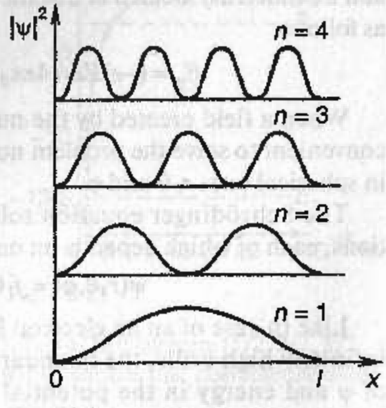


Fig. 28.9

Squaring (28.20), we will obtain probability density $|\psi|^2$ of the electron located different points of the potential well. Fig. 28.9 shows a curve of dependence of $|\psi|^2$ on x at different discrete states, i.e. at different quantum numbers. As we can see in the figure, an electron can be in different places of the potential well with different probability. There are such points, where the probability of finding

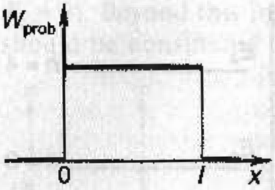


Fig. 28.10

electron is zero. It differs from the ideas of classical physics very much according to which, a particle can be found with equal probability in different places of the potential well (fig. 28.10), and it is impossible to divide the well with points where there can be no particle.

The Schrödinger equation can also be applied to more complicated fields of force, for examp to an electron in an atom. This will cause additional mathematical difficulti but will not change the main features of atomic systems: discreteness of ene states, probabilistic judgment about the location of an electron, unique |ψ dependence on coordinates, etc.

§ 28.6. APPLICATION OF THE SCHRÖDINGER EQUATION TO A HYDROGEN ATOM. QUANTUM NUMBERS

Description of atomic and molecular states using the Schrödinger equation is a quite difficult task. It is solved most simply for one electron located in the field of the nucleus. Such systems correspond to a hydrogen atom and hydrogen-like ions (helium atom once ionized, lithium atom twice ionized, etc.). However, in this case, solving this problem is beyond our course, so we will limit ourselves to a qualitative exposition of the problem.

First of all, into the Schrödinger equation, we should substitute potenti energy (28.13), which for two interacting point charges, which are e (electro and Ze (nucleus) located at distance r from each other in vacuum, is express as follows:

$$E_p = (-e)Ze/(4\pi\varepsilon_0 r) = -Ze^2/(4\pi\varepsilon_0 r). \qquad (28.22$$

When a field created by the nucleus has a center of symmetry, it is more convenient to solve the problem not in Cartesian rectangular coordinates, but in spherical ones r, θ and φ.

The Schrödinger equation solution is found as a product of three functions, each of which depends on one variable:

$$\psi(r,\theta,\varphi) = f_1(r)f_2(\theta)f_3(\varphi). \qquad (28.23$$

Like in case of an an electron located in a rectangular potential well with infinitely high walls, the boundary conditions led to specific possible valu of ψ and energy in the potential well corresponding to a hydrogen atom the physical conditions have led to possible values of f_1, f_2, f_3 and, therefo of ψ-function too. In this, discreteness of states becomes manifest too which is the main distinctive feature of quantum mechanical systems.

From the mathematical viewpoint, discreteness means that any of the functions of equation (28.23) has a set (spectrum) of solutions, each of which corresponds to a certain quantum number. In contradistinction to a rectangular potential well with infinitely high walls, the state of an electron in an atom is characterized not by one, but by several quantum numbers[1]. The first of them is the ***principal quantum number n***; it is equal to 1, 2, 3... It determines the energy levels of the electron according to the following law:

$$E = -me^4Z^2/(8\varepsilon_0^2h^2n^2). \qquad (28.24)$$

This expression is the solution to the Schrödinger equation and completely coincides with the corresponding formula of Bohr's theory (see § 28.7).

Fig. 28.11 shows the levels of possible values of the hydrogen atom total energy (E_1, E_2, E_3, etc.) and the curve of dependence of potential energy E_p on distance r between an electron and the nucleus [see (28.22)]. Upon an increase in principal quantum number n, r increases [see, for example, (28.33)], and the total [see (28.24)] and potential energy tends to zero. The kinetic energy also tends to zero. The shaded area ($E>0$) corresponds to the free electron state.

The second quantum number is an ***orbital quantum*** number l, which at given n can take values 0, 1, 2,.., $n - 1$. This number characterizes orbital momentum L of an electron relative to the nucleus:

$$L_l = \frac{h}{2\pi}\sqrt{l(l+1)}. \qquad (28.25)$$

The third quantum number is the *magnetic number ml*, which, at given l, takes values 0, ±1, ±2 ... ±l; $2l + 1$ values in total. This number determines the projection of the orbital momentum of the electron onto a certain direction Z arbitrarily chosen:

$$L_{lz} = \frac{h}{2\pi}m_l. \qquad (28.26)$$

The fourth quantum number is ***spin*** (***magnetic spin***) number[2] m_s.

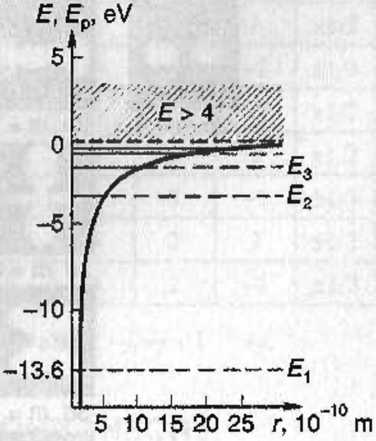


Fig. 28.11

---

[1] In general, quantum numbers are integers (0, 1, 2...) or half-integers (1/2, 3/2, 5/2...) which determine possible discrete values of physical quantities that characterize quantum systems and elementary particles.

[2] The presence of spin in particles does not follow from the Schrödinger equation.

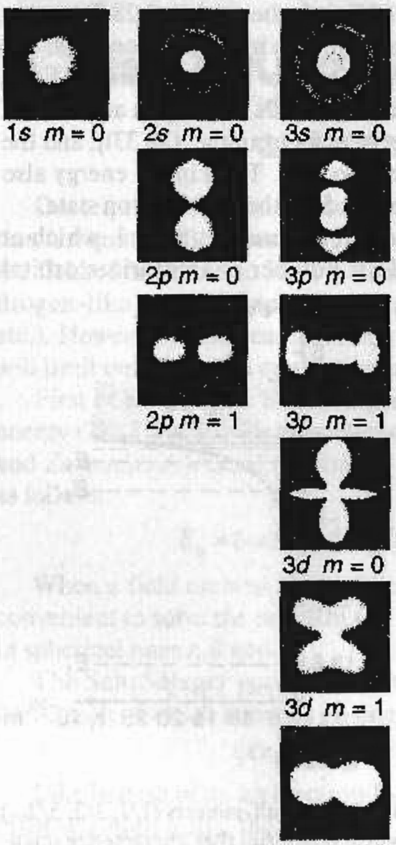
It can take only two values (±1/2) and characterizes possible values of sp projection of an electron:

$$L_{sz} = \frac{h}{2\pi} m_s. \qquad (28.2$$

The state of an electron in an atom with n and l given is designated in th following way: 1*s*, 2*s*, 2*p*, 3*s*, etc. In this, the digit means the value of the principal quantum number, and the letter is for the orbital quantum number: symbols *s, p, d, f*... are correspond by values $l = 0, 1, 2, 3$, etc.

The number of states with n and l given will be $2(2l + 1)$. To find the total number of states with the same principal quantum number, we sum up $2(2l + 1)$ for all possible values of l:

$$\sum_{l=0}^{l=n-1} 2(2l+1) = 2n^2. \qquad (28.2$$



1*s* $m = 0$ 2*s* $m = 0$ 3*s* $m = 0$

2*p* $m = 0$ 3*p* $m = 0$

2*p* $m = 1$ 3*p* $m = 1$

3*d* $m = 0$

3*d* $m = 1$

3*d* $m = 2$

Fig. 28.12

Thus, the first energy level of the hydrogen atom is corresponded by two states of the electron, the second level is done by 8 states, the third level is done by 18 states, etc. (Table 28.1).

In an atom, a visual representati of the electron location is obatined using electron cloud photograp (fig. 28.12). The photos were taken using a model with a lighting bulb. Havin calculated probability densities $|\psi|$ of an electron location in an atom, the bulb was moved in accordance with this calculation: it stayed longer in places with higher probability density, less time in places with a lower probabili density. Because of exposure, the photo film shows places of different intensit such places illustrate the distribution of the electron in an atom. The figure show how in relation to the moveme of the electron, the concept of an orbit is conventional and even misleading.

Spin and orbital magnetic momentums interact with each other, which changes the system of energy levels of an atom in comparison with what it would have been without such interaction. They say that spin-orbital interaction results in a fine structure of energy levels. If the interaction is essential, it is necessary to take into account the total angular momentum of the electron, i.e. orbital momentum and spin one. In this case, instead of m_l and m_s, other quantum numbers are used: j and m_j.

Quantum number j, which is the sum of the orbital number and spin, determines discrete values of total angular momentum L of an electron:

$$L = \frac{h}{2\pi}\sqrt{j(j+1)}. \tag{28.29}$$

Magnetic quantum number m_j characterizes possible projections of the total angular momentum onto a certain arbitrarily chosen direction Z:

$$L_Z = \frac{h}{2\pi}m_j. \tag{28.30}$$

At given l, quantum number j takes two values: ±1/2 (Table 28.1).

Table 28.1

State symbol	Quantum number values				State symbol	Quantum number values			
	n	l	m_l	m_s		n	L	m_l	m_s
1s	1	0	0	±1/2	3p	3	1	1	±1/2
2s	2	0	0	±1/2	3p	3	1	–1	±1/2
2p	2	1	0	±1/2	3d	3	2	0	±1/2
2p	2	1	1	±1/2	3d	3	2	1	±1/2
2p	2	1	–1	±1/2	3d	3	2	–1	±1/2
3s	3	0	0	±1/2	3d	3	2	2	±1/2
3p	3	1	0	±1/2	3d	3	2	–2	±1/2

At given j, quantum number n_{ij} takes $2j + 1$ values: $-j, -j + 1 \ldots +j$.

§ 28.7. NOTION OF BOHR THEORY

As early as in 1913 before quantum mechnaics was developed, a Danish physicist N. Bohr put forward the theory of the hydrogen atom and hydrogen-like ions, which was based on a nuclear model of the atom and its two postulates. Bohr's postulates did not fit in the framework of classical physics.

According to the first postulate, an atom and atomic systems can be only in certain stationary states for a long time. When in such states, the atom do not emit or absorb energy. Discrete energy values correspond to stationary states: E_1, E_2...

Any change in the energy of an atom or an atomic system is connect with an abrupt transition from one stationary state to another.

According to the second postulate, when an atom transits of from one state to another, the atom emits or absorbs a photon, whose energy is determin by equation (29.1).

Transition from a higher energy state to a lower energy state is accompanied by emission of a photon. The reverse process is possible with phot absorption.

According to Bohr's theory, in a hydrogen atom, an electron moves in a circular orbit around the nucleus. Of all possible orbits, stationary states correspo only to those at which the angular momentum is equal to integer quantity $h/(2\pi$

$$m\upsilon_n r_n = \frac{h}{2\pi} n \quad (n = 1, 2, 3 \ldots), \tag{28.31}$$

where m is the electron mass; υ_n is its velocity in the n-th orbit; r_n is the orbit radius. An electron, rotating in a circular orbit in an atom, is affected by the Coulomb force of attraction from a positively charged nucleus, which, accord to Newton's second law, is equal to the product of the mass of the electro by centripetal acceleration (the formula is valid for vacuum):

$$\frac{Ze \cdot e}{4\pi\varepsilon_0 r_n^2} = \frac{Ze^2}{4\pi\varepsilon_0 r_n^2} = \frac{m\upsilon_n^2}{r_n}, \tag{28.32}$$

where e is the electron charge; Ze is the nucleus charge. For hydrogen, $Z = 1$, for hydrogen-like ions, $Z > 1$. Eliminating υ_n from (28.31) and (28.32), we obtain

$$r_n = \varepsilon_0 h^2 n^2 / (\pi Z e^2 m). \tag{28.33}$$

Using (28.32), we find the electron kinetic energy:

$$E_k = m\upsilon_n^2/2 = -Ze^2/(8\pi\varepsilon_0 r_n), \tag{28.34}$$

and the sum of kinetic (28.34) and potential (28.22) energies yields the total electron energy:

$$E = E_k + E_p = \frac{Ze^2}{8\pi\varepsilon_0 r_n} - \frac{Ze^2}{4\pi\varepsilon_0 r_n} = -\frac{Ze^2}{8\pi\varepsilon_0 r_n}. \tag{28.35}$$

Substituting expression (28.33) into (28.35), we obtain [see (28.24)]:

$$E = -me^4 Z^2 / (8{\varepsilon_0}^2 h^2 n^2).$$

Based on the second postulate [see (29.1)] and formula (28.24), Bohr obtained formula (29.2), which explains serial regularities of the spectrum of the hydrogen atom and hydrogen-like ions (see § 29.2).

In Bohr's time, his theory was a triumph of atomic physics development. For the first time, although in a simplest atomic system (one electron moving around the nucleus), the regularities of the spectra were discovered.

Despite the great success of Bohr's theory, its drawbacks soon became noticeable. Thus, in the framework of this theory, it was not possible to explain the difference in the intensities of spectral lines, that is, to answer the question why some energy transitions are more probable than other. Bohr theory failed to reveal spectral regularities of a more complicated atomic system — a helium atom (two electrons orbiting the nucleus).

Another drawback of Bohr's theory was its inconsistency. This theory was neither classical nor quantum; it combined concepts of fundamentally different theories: classical and quantum physics. Thus, for example, Bohr's theory states that an electron moves in an atom in a definite orbit (classical concepts), but it does not emit an electromagnetic wave (quantum concepts).

In the first quarter of the 20th century, it became clear that Bohr's theory should be replaced with another theory of the atom. Quantum mechanics began to develop.

§ 28.8. ELECTRON SHELLS OF COMPLEX ATOMS

Quantum numbers describing an electron actual state in a hydrogen atom are used for approximate description of individual electrons states in complex atoms. However, at least two sizeable differences between complex atoms and hydrogen atom should be taken into account:

1) in complex atoms, the energy of electrons depends not only on n, but also on l due to their interaction;
2) the difference is due to the Pauli principle, according to which, an atom cannot have two (or more) electrons with four equal quantum numbers.

When an electron configuration corresponding to the normal state is formed, each electron of the atom tends to have the lowest energy. But for the Pauli principle, all electrons would be positioned at the lowest energy level. Actually, with certain exceptions, electrons have the same sequence of states that is given for a hydrogen atom in Table 28.1.

Electrons with an equal principal quantum number form a subshell. The subshells are designated K, L, M, N and so on in accordance with

$n = 1, 2, 3, 4$... Electrons with equal pairs of values of n and l are part of the subshell, which is briefly designated like the corresponding states for the electr of a hydrogen atom: $1s$, $2s$, $2p$, etc. Thus, for example, $2s$ subshell and $2s$ electron, etc.

In the subshell, the number of electrons is denoted as a superscript of the symbolic designation of the shell, for example, $2p^4$.

In an atom, electron distribution across subshells (electron configuratio is usually indicated as follows: for nitrogen $1s^2$, $2s^2$, $2p^3$, for calcium $1s^2$, $2s^2$, $2p^6$, $3s^2$, $3p^6$, $4s^2$, etc.

Since the electron energy of complex atoms depends not only on n, but also on l, formation of the Mendeleev table does not always take place by gradu filling of the subshells as the atom structure becomes more complicat For potassium ($Z = 19$), for example, instead of filling subshell M (possibl it was $1s^2$, $2s^2$, $2p^6$, $3s^2$, $3p^6$, $3a^1$), filling of subshell N begins, and the following electron configuration is formed: $1s^2$, $2s^2$, $2p^6$, $3s^2$, $3p^6$, $4s^1$.

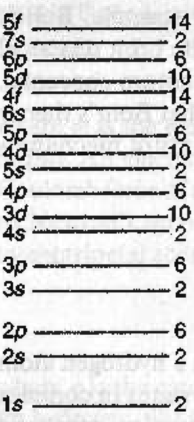


Fig. 28.13

Other elements have similar deviations from regul subshell filling.

The following general rule is always observed: the electrons of an unexcited atom are in a state with the lowe energy and in accordance with the Pauli princip Fig. 28.13 (not to scale) gives a schematic representat of energy states of a complex atom and the correspond number of electrons.

In conclusion, we will note that on the whole, the state of a multi-electron atom is determined by the followin quantum numbers. L is the total orbital momentum of the atom; L takes the values of 0, 1, 2, 3, etc.[1]; J is the total momentum of the atom; J can take values with the interval of 1 from $|L - S|$ to $|L + S|$; S is the resulting spin momentum of the atom; magnetic mJ that determines discrete values of the atom total momentu projected onto a certain axis Z:

$$L_{Az} = \frac{h}{2\pi} m_J. \qquad (28.3$$

At given J, m_J takes $2J + 1$ values:

$$-J, -J + 1 \ldots +J.$$

---

[1] Do not confuse this designation with the name of electron shell L and with the total angular momentum of an electron.

§ 28.9. ENERGY LEVELS OF MOLECULES

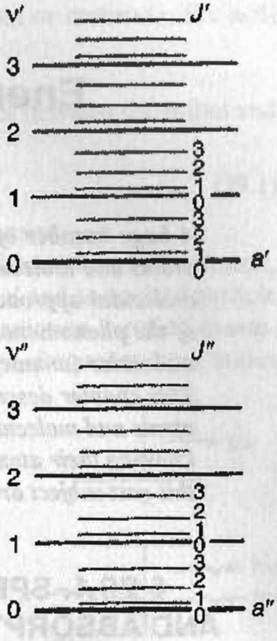


Fig. 28.14

Since molecules consist of atoms, motion inside molecules is more complicated than motion inside atoms. In a molecule, in addition to the motion of electrons around the nucleus, there is oscillatory motion of atoms around their equilibrium positions (oscillation of nucleus together with the electrons surrounding them) and rotational motion of the molecule as a whole.

Three types of energy levels correspond to electronic, oscillative, and rotational motion of the molecule: E_{el}, E_{osc} and E_{rot}. In a molecule, according to quantum mechanics, the energy of all types of motion takes only discrete values (it is quantized). Let us approximately represent total energy E of a molecule as a sum of quantized values of different energy types:

$$E = E_{\mathrm{el}} + E_{\mathrm{osc}} + E_{\mathrm{rot}}. \qquad (28.37)$$

Fig. 28.14 gives a schematic representation of the system of molecular levels: remote electron energy levels a' and a'', for which $E_{\mathrm{osc}} = E_{\mathrm{rot}} = 0$; close oscillation levels v', v'', for them $E_{\mathrm{rot}} = 0$; the closest rotational levels J' and J'' with different values of E_{rot}.

The difference between electron levels is in the region of several electronvolt; between adjacent oscillation levels, it is 10^{-2}–10^{-1} eV, between adjacent rotational levels, it is 10^{-5}–10^{-3} eV.

Chapter 29

Energy emission and absorption by atoms and molecules

A huge number of different phenomena are due to changing energy of atoms and molecules. In some cases, there is practical need for atomic-molecular approach to analysis of phenomena. In other cases, effective use of the phenomenon is possible only with compulsory consideration of its molecular (atomic) nature.
This chapter describes the specifics of energy emission and absorption by atoms and molecules, as well as certain practically important phenomena knowing their atomic nature is essential for their use. Certain problems of this vast subject are discussed in the next chapter.

§ 29.1. SPECIFICS OF ENERGY EMISSION AND ABSORPTION BY ATOMS AND MOLECULES

An atom and molecule can be in stationary energy states. In these states they do not emit nor absorb energy. Energy states are sketchily represented in the form of levels (see, for example, fig. 28.13). The lowest energy level — the basic one — corresponds to the ground state.

Upon quantum transitions, atoms and molecules jump from one stationary state to another, from one energy level to another.

A change in the state of atoms is connected with energy transitions of electrons. In molecules, energy can change not only because of electron transitions, but also due to changes in oscillations of atoms and transitions between rotational levels.

For transition from higher energy levels to lower ones, an atom or a molecule emits energy; in reverse transitions they absorb it. An atom in the ground state can only absorb energy.

There are the following two types of quantum transitions.

1. Without emission or absorption of electromagnetic energy by an atom or molecule: such as non-radiative transition occurs when an atom or a molecule interacts with other particles, for example, in a process of collision. There is inelastic collision, in which the internal state of the atom changes, and a non-radiative transition occurs, and elastic one

with a change in the kinetic energy of the atom or molecule, but with conservation of the internal state.

2. A photon is emitted or absorbed.

The photon energy is equal to the energy difference between the initial and final stationary states of an atom or molecule:

$$h\nu = E_i - E_k. \qquad (29.1)$$

Formula (29.1) expresses the ***law of energy conservation***.

Depending on the cause of quantum transition with emission of a photon, there are two types of emission. If this cause is internal and an excited particle transits to a lower energy level spontaneously, such radiation is called *spontaneous* (fig. 29.1, a). It is random and chaotic as for time, frequency (transitions between different sublevels are possible), direction of propagation and polarization. Conventional light sources emit mostly spontaneously. The other type of radiation is ***forced*** or ***stimulated*** (fig. 29.1, b). It occurs when a photon interacts with an excited particle, if the energy of the photon is equal to the difference in energy levels. Due to forced quantum transition, two identical photons will propagate from the particle in the same direction: one is primary, forcing, and the other is secondary, emitted.

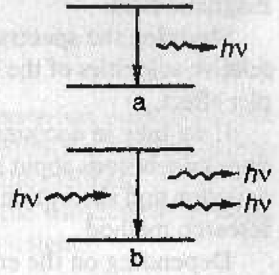


Fig. 29.1

Energy emitted by atoms or molecules forms the emission spectrum, while absorbed energy forms the absorption spectrum.

Intensity of spectral lines is determined by the number of identical transitions occurring per second, and therefore, it depends on the number of emitting (absorbing) atoms and probability of the corresponding transition.

Quantum transitions do not occur between any energy levels. There are rules of selection, or prohibition, forming conditions, at which transitions are possible, impossible or almost impossible.

The energy levels of most atoms and molecules are quite complicated. The structure of levels, and therefore, of spectra, depends not only on the structure of a certain atom or molecule, but also on external conditions.

Electromagnetic interaction of electrons leads to a fine splitting[1] of energy levels (the fine structure). The influence of magnetic moments of nuclei causes hyperfine splitting (the super-fine structure). Electric and magnetic fields, which are beyond an atom or molecule also make energy levels to divide (Stark and Zeeman phenomena; see § 30.2).

---

[1] Here the term "splitting" does means a state, which has already occurred, rather than a process.

Spectra are a source of different information.

First, the type of spectrum can identify atoms and molecules, which is a part of objectives of qualitative spectral analysis. The intensity of spectral lines determines the number of emitting (absorbing) atoms; it is used in quantitative spectral analysis. It is quite easy to detect impurities in concentration of 10^{-5}–10^{-6}% and determine the composition of samples with very low mass down to several tens micrograms.

By spectra, it is possible to judge of the structure of atoms or molecul and their energy levels, mobility of separate parts of large molecules, etc. Knowing the dependence of the spectra on fields acting on an atom or molecule, we can learn about mutual arrangement of the particles, because the impact of adjacent atoms (molecules) is carried out by means of an electromagnetic field.

Studying the spectra of moving body makes it possible to determine the relative velocities of the emitter and radiation receiver based on optical Doppler effect.

If we take in account that the spectrum of a material makes it possible to draw conclusions about its actual state, temperature, pressure, etc., we see that emission and absorption of energy by atoms and molecules is a highly valuable research method.

Depending on the energy (frequency) of a photon emitted or absorbed by an atom (or molecule), the following spectroscopy methods are distinguish ***radio-*, *infrared*, *visual radiation*, *ultraviolet and X-ray***[1] spectrosocpy.

Depending on the type of a material (spectrum source), they distinguis atomic, molecular, and crystal spectra.

§ 29.2. LIGHT ABSORPTION

The intensity of light propagation in the medium can decrease due to absorption and scattering by molecules (atoms) of the substance.

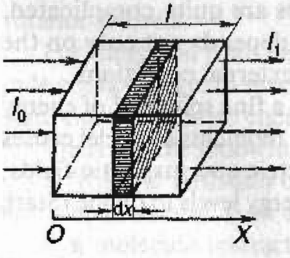


Fig. 29.2

Light absorption *is attenuation of light intensity when passing through any material due to conversion of light energy into other types of energy.*

Let us determine the law of light absorption by a material. If we select a thin layer of a material of thickness dx (fig. 29.2), then, in case of absorption, attenuation of light intensity dI by thi layer will be the higher the more the thickness of the layer and the intensity of light incident on this layer are:

---

[1] Here gamma spectroscopy caused by nuclear quantum transitions is not indicated

$$\mathrm{d}I = -kI \cdot \mathrm{d}x, \tag{29.2}$$

where k is the natural absorption coefficient (proportionality coefficient, which depends on the absorbing medium and does not depend, within certain limits, on the intensity of light); the minus sign means that the intensity of light, as it passes through the material, decreases, thus $\mathrm{d}I<0$. Integrating (29.2) and substituting the correspondent limits (fig. 29.2), we obtain:

$$\int_{I_0}^{I_l}\frac{\mathrm{d}I}{I} = -k\int_0^l \mathrm{d}x \quad \text{or} \quad \ln\frac{I_l}{I_0} = -k$$

from where, finding the antilogarithm, we have:

$$I_l = I_0 e^{-kl}. \tag{29.3}$$

This formula expresses ***Bouguer's law of light absorption***. As we can see, natural absorption coefficient k is reciprocal to the distance at which light intensity is attenuated e times because of absorption in the medium.

The natural absorption coefficient depends on the wavelength of light, so it is expedient to write law (29.3) for monochromatic light:

$$I_l = I_0 e^{-k_\lambda l}, \tag{29.4}$$

where k_λ is the monochromatic natural absorption coefficient.

Since light absorption is due to interaction with molecules, the law of absorption can be associated with certain characteristics of molecules.

Let n be the concentration of molecules absorbing quanta of light. Absorption cross section of a molecule we designate as σ (it is the area within which the photon is captured by the molecule).

If we consider that the cross section area of a rectangular parallelepiped (fig. 29.2) is equal to S, then the volume of the layer selected is $S\mathrm{d}x$, and the number of molecules in it is $nS\mathrm{d}x$.

The total area of absorption cross section of the molecules of this layer is $\sigma nS\mathrm{d}x$. The flow of photons incident on this layer is $\Phi = IS$. In the total cross-sectional area, the share of absorption cross section area of molecules is:

$$\frac{\sigma nS\mathrm{d}x}{S} = \sigma n\mathrm{d}x. \tag{29.5}$$

It can be considered that molecules absorb the the same share [see (29.5)] of photons incident on the layer, because the ratio between these areas determines the probability of one photon interacting with molecules of the layer selected.

The part of photons absorbed by the layer can be expressed through th flux ($d\Phi/\Phi$) or intensity (dI/I) of light. Based on the stated above, it is possi to write down:

$$\frac{dI}{I} = -\sigma n dx, \qquad \text{(29.[illegible])}$$

from where, after integration and finding the antilogarithm, we have:

$$I_l = I_0 e^{-\sigma n l}. \qquad (29.7)$$

This equation, in contradistinction to (29.4), includes the molec parameter σ.

Let us suppose that photon-absorbing molecules are in a solvent that do not absorb light. Molar concentration $C = n/N_A$, from which, $n = CN_A$.

Let us transform the product σn:

$$\sigma n = \sigma C N_A = \chi' C,$$

where $\chi' = \sigma \cdot N_A$ is natural molar absorption coefficient.

Its physical meaning is the total absorption cross section for all molec of one mole of dissolved material. Using this concept, it is possibl to write (29.7) in the following form (***Bouguer–Lambert–Beer* law**):

$$I_l = I_0 e^{-\chi' C l}. \qquad \text{(29.[illegible])}$$

In laboratory practice, the Bouguer–Lambert–Beer law is usually expressed through the exponential function with base 10:

$$I_l = I_0 \cdot 10^{-\chi' C l}, \qquad (29.9)$$

where χ is molar absorption coefficient; $\chi \approx 0.43\chi'$ since $e \approx 10^{0.43}$. Usually χ is ascribed to a certain wavelength and called monochromatic molar absorpti coefficient (χ_λ).

The ratio between the radiant flux passing through the given body or a solution and the corresponding incident flux is called ***transmittance***.

Let us express it as a ratio of intensities:

$$\tau = I_l / I_0. \qquad (29.10)$$

The common logarithm of a quantity is reciprocal of transmittan It called ***absorbance*** of a solution:

$$D = \lg(1/\tau) = \lg(I_0/I_l) = \chi_\lambda C l, \qquad (29.11)$$

Based on the Bouguer-Lambert-Beer law, a number of photometri methods of determining material concentration in a colored solution (***concen tion colorimetry***) are developed. These methods provide for direct meas ment of light flows that have passed through a solution, transmittance or absorbance.

The dependencies $k_\lambda = f_1(\lambda)$ and $\chi_\lambda' = f_2(\lambda)$ are spectra of material absorption.

Absorption spectra are sources of information about the actual state of a material and the structure of energy levels of atoms and molecules (see § 29.4 and 29.5).

§ 29.3. LIGHT SCATTERING

***Light scattering** is a phenomenon of light deflection in all possible directions when it propagates in a medium.*

A prerequisite for the occurrence of light scattering is the presence of optical inhomogeneities, that is areas with a refractive index which differs from the refractive index of the basic medium.

Scattering and diffraction of light have certain common features; both phenomena depend on the ratio of the obstacle or inhomogeneity dimensions to the wavelength. The difference between these phenomena is that diffraction is due to interference of secondary waves, while scattering is due to composition (rather than interference!) of radiation occurring because of forced oscillations of electrons in the inhomogeneities upon exposure to light.

There are two main types of inhomogeneities:

1. fine foreign particles in a homogeneous transparent material; such media are cloudy: smoke (solid particles in a gas), fog (liquid droplets in a gas), suspensions, emulsions, etc. Scattering in blurred media is called the Tyndall phenomenon;
2. optical inhomogeneities occurring in a pure material due to statistical deviation of molecules from uniform distribution center (density fluctuations). Light scattering on such inhomogeneities is called molecular scattering of light; for example, light scattering by atmosphere.

The light intensity attenuation due to dispersion, as well to absorption, is described using an exponential function:

$$I_l = I_0 e^{-ml}, \qquad (29.12)$$

where m is natural scattering index.

Under the joint impact of absorption and dispersion, attenuation of light intensity is also an exponential function:

$$I_l = I_0 e^{-\mu l}, \qquad (29.13)$$

where μ is natural attenuation index. It is easy to see that $\mu = m + k$.

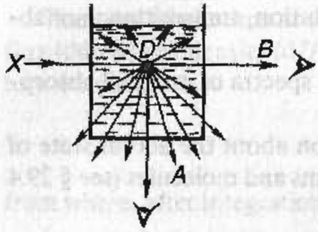


Fig. 29.3

Rayleigh determined that in a turbid medium with inhomogeneities approximately less than 0.2λ as well as upon molecular scattering the intensity of scattered light is inversely proportional to the fourth power of the wavelength (***Rayleigh's law***):

$$I \sim 1/\lambda^4. \qquad (29.14)$$

This means that white light, for example at point D (fig. 29.3), mainly light blue and violet rays will be scattered (direction A), and red rays will propagate in direction B of the incident light. A similar phenomenon is observed in nature: the light blue color of the sky is scattered light, the red color of the setting sun is a change in the spectrum of white light due to substantial scattering of blue and violet rays in the atmosphere at oblique incidence (see the explanation for fig. 27.3).

Lesser scattering of red rays is used in signaling: identification lights on airfields; the most important light of traffic lights is red, etc. Infrared rays are scattered even less. Fig. 29.4 shows two photos of the landscape: the left one was taken in a regular way; fog limited visibility in it very much; but in the right photo, taken in infrared radiation on a special plate, the fog does not interfere, it is transparent for longer wavelengths.

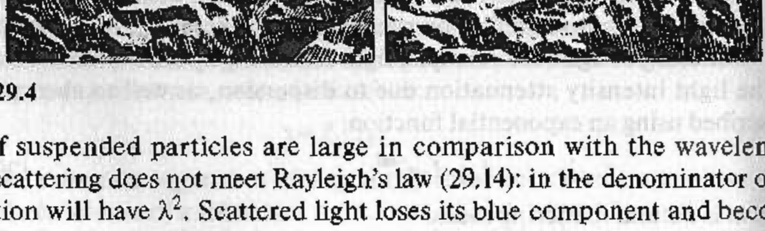


Fig. 29.4

If suspended particles are large in comparison with the wavelength the scattering does not meet Rayleigh's law (29.14): in the denominator of the fraction will have λ^2. Scattered light loses its blue component and becomes whiter. So, the dusty sky of cities seems to be whitish in contrast to the dark blue sky of clear expanses of the sea.

The direction of scattered of light, degree of its polarization, spectral composition, etc. inform us about parameters characterizing intermolecular interaction, sizes of macromolecules in solutions, particles in colloidal solutions emulsions, aerosols, etc. Methods of measuring scattered light to obtain such data are called *nephelometry*, and corresponding devices are called *nephelometers.*

§ 29.4. OPTICAL ATOMIC SPECTRA

***Atomic spectra** are spectra of both emission and absorption that occur upon quantum transitions between levels of free or weakly interacting atoms.*

By optical atomic spectra, we mean those that are due to transitions between the levels of outer electrons with photon energy of about several electron-volts. These are ultraviolet, visible and near infrared (of the order of micrometers) spectral regions.

The most interesting are optical atomic emission spectra from excited atoms. Their excitation is usually achieved due to non-radiative quantum transitions upon an electric discharge in a gas or heating a material with gas burner flame, electric arc, or spark.

In § 29.1 we discussed universal considerations of atomic spectra. Detailed information of specific atomic spectra can be found in reference books on spectroscopy. As a simple example, let us consider the spectrum of a hydrogen atom and hydrogen-like ions.

From formulas (28.24) and (29.1), it is possible to obtain the formula for the frequency of light emitted (absorbed) by a hydrogen atom ($Z = 1$):

$$\nu = \frac{me^4}{8\varepsilon_0^2 h^2}\left[\frac{1}{k^2} - \frac{1}{i^2}\right]. \qquad (29.15)$$

This formula was experimentally discovered by J.J. Balmer long before quantum mechanics was known and was theoretically obtained by Bohr (see § 28.7); *i* and *k* are ordinal numbers of the levels between which the quantum transition occurs.

In a spectrum, groups of lines called *spectral series* can be distinguished.

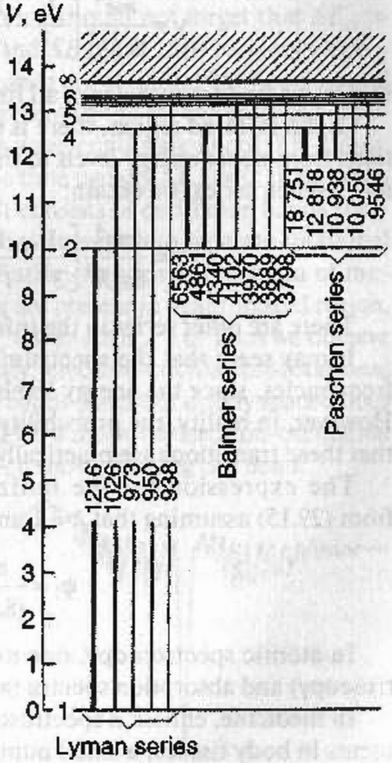


Fig. 29.5

With regard to emission spectra each series corresponds to transitions from different levels to the same final level (fig. 29.5).

In the ultraviolet region, there is the *Lyman series* formed by transiti from upper energy levels to the lowest, which is ground level ($k = 1$). From formula (29.15), for the Lyman series, we obtain:

$$\nu = \frac{me^4}{8\varepsilon_0^2 h^3}\left[\frac{1}{1^2} - \frac{1}{i^2}\right] \; (i = 2,3,4\ldots), \qquad (29.16)$$

that is, we find frequencies of all lines in this series. The longest wavelen line has the highest intensity. In fig. 29.5, spectral line intensities are conve tionally expressed with thickness of lines corresponding to transitions.

In the visible and near-ultraviolet regions of the spectrum, there is th Balmer series that occurs because of transitions from upper energy levels to the second level ($k = 2$). From formula (29.15), for *Balmer series,* we obtain:

$$\nu = \frac{me^4}{8\varepsilon_0^2 h^3}\left[\frac{1}{2^2} - \frac{1}{i^2}\right] \; (i = 3,4,5\ldots), \qquad (29.17)$$

that is, we find frequencies of all lines in this series.

In the infrared region, there is the *Paschen series* that occurs upon tran tions from upper energy levels to the third level ($k = 3$). From formula (29.15), for Paschen series, we obtain:

$$\nu = \frac{me^4}{8\varepsilon_0^2 h^3}\left[\frac{1}{3^2} - \frac{1}{i^2}\right] \; (i = 4,5,6\ldots). \qquad (29.18)$$

There are other series in the infrared region.

It may seem that the spectrum of atomic hydrogen is not limited by low frequencies, since the energy levels, as n increases, become arbitrarily clos However, in reality, the probability of transition between such levels is so low that these transitions are practically never seen.

The expression for the ionizing potential (15.15) can be obtain from (29.15) assuming that $k = 1$ and $i \to \infty$:

$$\varphi_i = \frac{me^4}{(8\varepsilon_0^2 h^2)}. \qquad (29.19)$$

In atomic spectroscopy, one uses both emission spectra (emission spectroscopy) and absorption spectra (atomic absorption spectroscopy).

In medicine, emission spectroscopy is mainly used to determine micro ments in body tissues, a small number of metal atoms in canned foods for sanitary purposes, certain elements in cadaver tissues for forensic purpose etc.

§ 29.5. MOLECULAR SPECTRA

Molecular spectra of emission and absorption occur upon quantum transitions of molecules from one energy level to another (see § 28.9) and consist of a set of more or less wide bands, that are closely spaced lines. In comparison with atomic spectra, the complexity of molecular spectra is due to a wide variety of motion types and, therefore, energy transitions in the molecule.

Taking into account (29.1) and (28.37), we find the frequency emitted or absorbed by a molecule:

$$\nu = \frac{1}{h}(E' - E'') = \frac{1}{h}\left[(E'_{\text{el}} - E''_{\text{el}}) + (E'_{\text{col}} + E''_{\text{col}}) + (E'_{\text{bp}} - E''_{\text{bp}})\right] = \\ = \frac{1}{h}(\Delta E_{\text{el}} + \Delta E_{\text{col}} + \Delta E_{\text{bp}}). \tag{29.20}$$

In this, one or two accents refer, as it is accepted in molecular spectroscopy, to the upper and lower levels respectively. One should not forget that $\Delta E_{\text{el}} \gg \gg \Delta E_{\text{osc}} \gg \Delta E_{\text{rot}}$. If $\Delta E_{\text{el}} = 0$ and $\Delta E_{\text{osc}} = 0$, and $\Delta E_{\text{rot}} \neq 0$, then molecular spectra consisting of individual pure rotational bands are obtained; they correspond to low frequencies [see (29.20)]. They are present in the far infrared and especially in the microwave (UHF) region. The wavelengths are about 0.1–1 mm.

If $\Delta E_{\text{el}} = 0$ and $\Delta E_{\text{osc}} \neq 0$, then at the same time usually $\Delta E_{\text{rot}} \neq 0$; in this case an oscillation-rotational spectrum occurs. It consists of oscillation bands splitting at the sufficient resolution of a spectral instrument into separate rotational lines. Fig. 29.6 graphically shows the oscillation-rotational spectrum of methane. Infrared oscillation-rotational spectra are present in near infrared region.

If $\Delta E_{\text{el}} \neq 0$ usually at the same time $\Delta E_{\text{osc}} \neq 0$ and $\Delta E_{\text{rot}} \neq 0$. Thus we observe electron oscillation spectra, or, more exactly, electron-oscillation-rotational spectra that consist of various bands, and the bands consist of closely spaced lines corresponding to rotational transitions. Fig. 29.7, a shows the electron-oscillation spectrum of a nitrogen molecule and fig. 29.7, b shows splitting of a band.

The electron-oscillation-rotational spectra of emission and absorption are present in the visual and ultra-violet region.

Specificity and individuality of spectra of individual molecules is the basis of qualitative an quantitative spectral analysis. Molecular spectra make it possible to study not only the structure of molecules, but also the nature of intermolecular interactions.

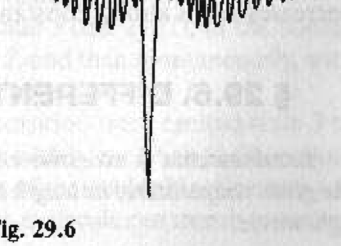


Fig. 29.6

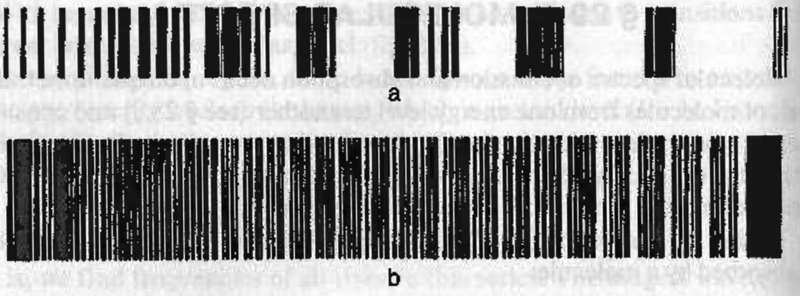


Fig. 29.7

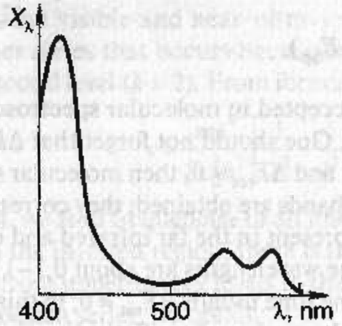


Fig. 29.8

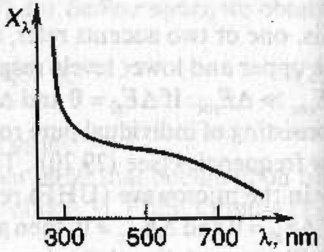


Fig. 29.9

Molecular absorption spectra are an important source of information about biologically functional molecules; they are widely used in modern biochemical and biophysical studies.

In many cases, these spectra are recorded as continuous ones without resolution of the details above.

For example, fig. 29.8 shows the absorption spectrum of erythrocyte suspension. The absorption spectrum of human skin is shown in fig. 29.9; in the ultraviolet region, the absorption coefficient is high and the skin absorbs radiation with its top layers. In the visible region, the absorption coefficient decreases and is almost constant up to the red region.

§ 29.6. DIFFERENT TYPES OF LUMINESCENCE

***Luminescence** is excessive emission by a body exceeding thermal emission at the given temperature; its length is much more than the period* (10^{-15} s) *of emitted light waves.*

In this definition, the length indicator was proposed by S.I. Vavilov to distinguish luminescence from certain other phenomena of afterglow, such as reflection and scattering of light.

Depending on the excitation type, there are several types of luminescence.

Luminescence caused by charged particles: ions — *ions luminescence*, electrons — *cathodoluminescence*, nuclear radiation — *radioluminescence*. Luminescence due to of X-ray and gamma radiation is called X-ray luminescence, luminescence due to photons is called *photoluminescence* (see § 29.7). Upon rubbing, crushing or cracking of certain crystals *triboluminescence* occurs. The electric field excites *electroluminescence*, its particular case is glowing of a gas discharge. Luminescence accompanying an exothermic chemical reaction is called *chemiluminescence* (see § 29.8).

§ 29.7. PHOTOLUMINESCENCE

***Photoluminescence**, called sometimes simply luminescence, is subdivided into fluorescence (short-term afterglow) and phosphorescence (relatively term afterglow).*

The initial act of any photoluminescence is excitation by a photon with energy $h\nu$ of an atom or molecule. In the simplest case, that usually takes place in single-atom vapors and gases, an atom returns to its ground state emitting a photon of the same frequency ν (fig. 29.10). This phenomenon is called *resonance fluorescence* (*resonance scattering*). Special experiments showed that such glow occurs in about 10^{-8} s after lighting of the material and, therefore, is not scattering in the usual sense of the word.

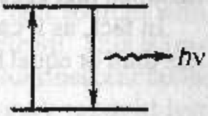


Fig. 29.10

When foreign gases (hydrogen, oxygen, etc.) are added into luminescent vapors, resonance fluorescence decreases. This is due to the fact that during the time the atom is in the excited state it can encounter a molecule of another kind and transfer energy to it. The kinetic energy of the molecule increases, and the atom goes into the ground state in a nonradiating way.

It is more probable that from the excited state *3* (fig. 29.11), in the nonradiating way, the molecule will move onto level *2*, and then spontaneously, with emission of a quantum with energy $h\nu'$, it will move onto level *1*.

In complex organic molecules, there is transition from excited state 3 to a certain intermediate metastable state *4*, from which transition to the ground state is improbable (fig. 29.12). At the expense of molecular kinetic energy of the surrounding particles or a new quantum, the molecule can transit to excited

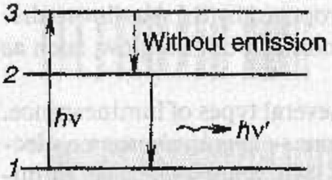


Fig. 29.11

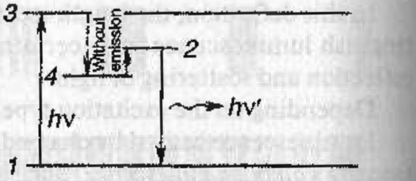


Fig. 29.12

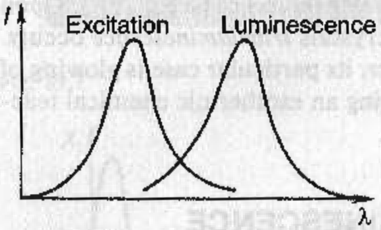


Fig. 29.13

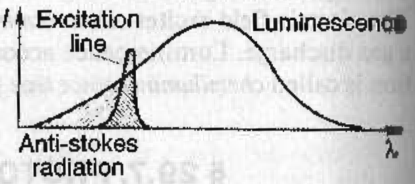


Fig. 29.14

level *2* and from it to ground state *1*. This is one of the mechanisms of phosphorescence. Heating increases the probability of leaving the metastable level and increases phosphorescence.

In photoluminescence, *Stokes law* is generally valid; **a luminescence spectrum is shifted in the direction of long wavelength relative to the spectrum of radiation that caused it** (fig. 29.13).

In fact, as it can be seen in fig. 29.10, energy $h\nu'$ of the emitted photon is less than or equal to energy $h\nu$ of the absorbed photon:

$$h\nu' \le h\nu, \qquad (29.21)$$

whence $\lambda' \ge \lambda$. There are deviations from the Stokes law, which are called anti-Stokes luminescence. This is seen especially well when photoluminescence is excited by an individual spectral line, i.e. by monochromatic light (fig. 29.14). Anti-Stokes emission occurs upon excitation of a particle, which has already been in an excited state (fig. 29.15, level *3*). Upon transition from level *2* onto ground level *1*, energy $h\nu'$ is emitted. As we can see from the figure:

$$h\nu' > h\nu \quad \text{or} \quad \lambda' < \lambda. \qquad (29.22)$$

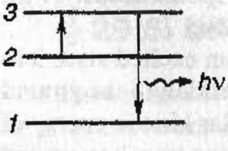


Fig. 29.15

A number of biologically functional molecules, such as protein molecules, have fluorescence. Fluorescence factors are sensitive to the structure of the fluorescent molecule environment; therefore, luminescence can be used to study chemical transformations and intermolecular interaction.

In recent decades, special fluorescent molecules added to membrane systems externally have become widely used. Such molecules are called fluorescent probes (non-covalent bond with the membrane) or fluorescent labels (chemical bond).

Changes in the fluorescence of probes and labels make it possible to detect conformational rearrangements in proteins and membranes.

Let us consider certain applications of photoluminescence for medical purposes.

Luminescence analysis based on observation of luminescence of objects for the purpose of their study is used to detect the initial stage of food spoilage, sorting pharmacological preparations, and diagnostics of some diseases. For instance, in ultraviolet light, hair and scales affected by fungus give a bright green luminescent glow. The permeability of skin capillaries can be determined injecting luminescent dyes subcutaneously.

Under favorable conditions, luminescent analysis makes it possible to detect luminescent materials with a mass down to 10^{-10} g.

Luminescent analysis of microscopic objects is carried out with special luminescent microscopes in which, in contrast to conventional light sources, high and ultrahigh pressure mercury bulbs are commonly used. Two light filters are used too. One of them is placed in front of the condenser, it separates the light source region which induces the object to luminesce; the other filter located between the objective lens and the eyepiece separates the light of luminescence. Based on photoluminescence, one can design light sources with a spectrum closer to daylight than incandescent bulb spectrum. This is important for both industrial and sanitary purposes. In such luminescent bulbs called daylight bulbs an electric discharge occurs in mercury vapor at low pressure (electroluminescence). Onto the inner surface of the bulb made from conventional glass (fig. 29.16) a thin layer of luminophore is applied, which generates photoluminescence under the impact of of mercury vapor radiation.

By changing the composition of the luminophore, it is possible to choose the most suitable photoluminescence spectrum. Fig. 29.17 shows a possible spectrum, the intense lines correspond to the spectrum of vapors whose radiation passes through the luminophore partially.

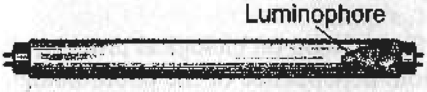


Fig. 29.16

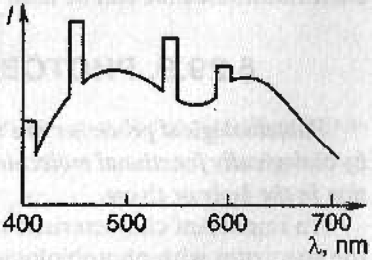


Fig. 29.17

§ 29.8. CHEMILUMINESCENCE

Luminescence accompanying chemical reactions is called chemiluminescence.

It is generated either directly by reaction products or by other components that are excited because of energy transfer to them from the reaction products.

Luminance of chemiluminescence, i.e. the number of quanta emitted per time unit, increases along with the rate of reaction and efficiency of chemiluminescence. Chemiluminescence is the average number of quanta per elementary reaction. Chemiluminescence makes it possible to determine the composition of a material (chemiluminescence analysis).

A particular manifestation of chemiluminescence is a glow accompanying chemical reactions in biological objects. It is called ***biochemical luminescence***. Glowing of rotten wood and fireflies are examples of biological chemiluminescence (bioluminescence).

Among biophysicists, low-intensity bioluminescence is called ***ultra weak glowing***; it was actively studied by a number of scientists, in particular, by Yu.A. Vladimirov.

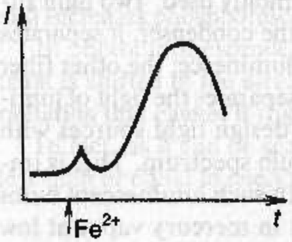


Fig. 29.18

It was shown that in biological systems, chemiluminescence occurs upon recombination of lipid peroxide free radicals: $RO_2 + RO_2 \rightarrow$ excited product $\rightarrow$ product + chemiluminescence quantum.

Intensity of chemiluminescence increases very much when, for example, bivalent iron salts are added to the biological systems being studied. Fig. 29.18 shows an increase in the luminescence intensity in a mitochondrial suspension at the time point of additing bivalent iron salts. If such an experiment is done with blood plasma in case of purulent appendicitis or cholecystitis, we can see that in the first case, the glow is much weaker. Thus, chemiluminescence can be used as a diagnostic method.

§ 29.9. PHOTOBIOLOGICAL PROCESSES

Photobiological processes are those that begin with absorption of light quanta by biologically functional molecules and end in corresponding physiological reaction in the body or tissue.

An important characteristic of the effect of light on biological processes is the spectrum with photobiological action: dependence of the photobiological effect on the wavelength of the active light. Action spectra make it possible to determine which region of the spectrum induces a biological process most effectively and to find out the mechanism of such actions.

A physician must understand these processes to explain the mechanism of vision (see § 29.10) and estimate various effects of UV-radiation (see § 27.7).

Absorbing a quantum of light (see § 29.2), a molecule gets excited. Its excitation energy can be transferred to other molecules. It is significant for a photobiological process that a chemical transformation (photochemical reaction) should take place due to this excitation. After the primary photochemical exposure, the reaction develops so that the presence of light is not necessary (dark reactions); after all, they lead to a response of the biological system to light.

Let us consider the initial stages of this process quantitatively: light absorption and primary photochemical reaction.

Similarly to § 29.2, let us introduce a notion of photon absorption cross section σ by a molecule. It differs from the derivation of the Bouguer–Lambert–Beer law, at least, the following: first, we will take into account the decrease in the number of activated molecules: action of light causes their transformation. Second, let us consider a sufficiently thin layer of a solution diluted; this permits regarding the intensity of light I_0 as constant and the same throughout the entire solution.

Elementary reduction of concentration dn of molecules upon exposure to light is proportional to:

- concentration n of molecules;
- absorption cross-section σ;
- radiation time dt;
- number of photons I_0 that pass through 1 m^2 of the cell face per time:

$$\mathrm{d}n = -\varphi_x I_0 \sigma n \,\mathrm{d}t. \qquad (29.23)$$

In this, the minus sign denotes a decrease in the number of molecules over time. Coefficient φ_x is called quantum yield of a photochemical reaction. This coefficient shows what portion of molecules that have absorbed photons enter the photochemical reaction. Let us separate the variables and integrate (29.23):

$$\int_{n_0}^{t} \frac{\mathrm{d}n}{n} = -\varphi_x I_0 \sigma \int_0^t \mathrm{d}t,$$

where n_0 is the initial molecule concentration and n_t is the molecule concentration at time point t. We obtain:

$$\ln(n_t/n_0) = -\varphi_x I_0 \sigma t,$$

or

$$\ln(n_0/n_t) = \varphi_x I_0 \sigma t = D_0 \sigma_x. \qquad (29.24)$$

In this, $I_0t = D_0$ is radiation exposure and $\sigma\varphi_x = \sigma_x$ is the cross section [illegible] of a molecule for photochemical transformation; the area is proportional to the probability of a photon interacting with a molecule, which results in a photochemical.

To find φ_x a dependence curve $\ln(n_0/n_t) = f(D_0)$ is plotted, and by the slope of the straight line [see (29.24)], this value is found (fig. 29.19).

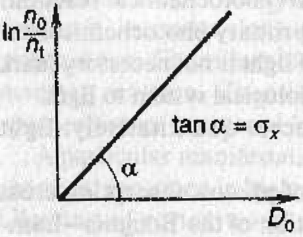


Fig. 29.19

In photochemistry, dependence $\sigma_x(\lambda)$ is called action spectrum. This dependence can be found using relation $\sigma_x = \sigma\varphi_x$. The point is that in solutions, the quantum yield of photochemical reactions does not depend on the wavelength of the acting light (φ_x = const). Physically, it means that irrespective of excitation energy $h\nu$ a molecule will be excited (see 28.9) and will be able to start its photochemical transformation. Taking this into account, we can conclude that action spectrum $\sigma_x(\lambda)$ and absorption spectrum (see § 29.2), i.e. dependence $\sigma(\lambda)$, have the same shape because they differ only by constant multiplier φ_x. This feature makes it possible, comparing the spectrum of photobiological action with the absorption spectra of different biochemical compounds, to determine the mechanism of action of light and, in particular, the mechanism of action of UV-radiation.

For example, it was determined that the curve of bacterial death induced by UV-radiation exposure (the spectrum of photobiological action) is like [illegible] absorption spectrum of nucleus acids. This led to the conclusion that death of bacteria is caused by damage exactly to nucleic acids.

§ 29.10. BIOPHYSICAL FUNDAMENTALS OF VISUAL PERCEPTION

In § 26.4, we considered the specifics of the light-conducting part of the eye. Perception of light by the eye is a photobiological process,so here we will consider the mechanism of the light-perceiving system.

Light-sensitive visual cells — rods and cones — play different roles in excitation of light sensation. Rods are more light-sensitive, but do not distinguish colors. Cones distinguish colors; in addition, at sufficient luminance of the object, they are sensitive to perception of image details, so the resolution of the eye is due to positioning of cones on the retina (see § 26.4).

The rods belong to the system of achromatic and twilight vision, and cones do to daylight and color vision.

Let us first consider certain general problems of the sensitivity of the eye to light and color.

Light perception by the eye is the reciprocal of the luminance threshold, i.e. the minimum luminance that causes visual sensation under actual conditions of observation.

Light perception by the eye varies widely due to visual *adaptation*, which is the ability of the eye to adapt to different brightness. Adaptation is carried out in the following ways.

1. Change in the pupil diameter in the range from 2 to 8 mm, which changes the luminous flux 16 times.
2. Reduction of the concentration of light-sensitive material.
3. Screening the cones and rods with a dark pigment, which is located in the vascular membrane and is able to move in the direction of the vitreous body in the process of adaptation.
4. Changing the proportion of rods and cones participating light sensation depending on the luminance of an object.

Adaptation makes it possible for the eye to function normally in a luminance range from 10^{-7} to 10^{5} cd/m^2. At full dark adaptation, the lower limit, or absolute threshold of eye light sensitivity is about a hundred photons per second. Only about 10% of them are absorbed by the molecules of visual pigment in the rods of the retina, and the rest are reflected from the cornea and absorbed by the optical media of the eye or pass through the retina and are absorbed by the cells of the pigment epithelium. The presence of retinal pigment epithelium lining the retina reduces reflection and scattering of light from the posterior wall of the eye. The human eye reacts to the electromagnetic waves with a wavelength from about 400 to 760 nm. Spectral sensitivity of the eye is characterized by the *luminous efficacy*:

$$S_\lambda = \mathrm{d}\Phi/\mathrm{d}\Phi_e, \tag{29.25}$$

where $\mathrm{d}\Phi$ is the luminous flux; $\mathrm{d}\Phi_e$ is the radiated power generating this luminous flux; $\mathrm{d}\Phi$ and $\mathrm{d}\Phi_e$ relates to the quite small interval of wavelengths from λ to $\lambda + \mathrm{d}\lambda$. It is accepted that the normal eye is the most sensitive to the yellow-green monochromatic radiation of $\lambda = 555$ nm. In this case, the radiated power, which is equal to 1 W, causes light perception corresponding to the luminous flux of 683 lm. From (29.25), this wavelength has the luminosity function $S_\lambda = 555\ \mathrm{nm} = 683$ lm/W.

A relative luminosity function is equal to

$$s_\lambda = S_\lambda / S_{\lambda\max}, \tag{29.26}$$

where $S_{\lambda\max}$ is the maximum luminosity function for this spectrum.

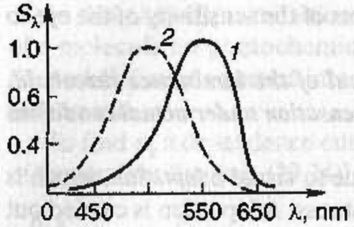


Fig. 29.20

Fig. 29.20 shows curve *1* of luminosity function of day vision and curve *2* of twilight vision *2*. In essence, these curves are the spectrum of photobiological action. For day vision, $S_{\lambda\max}$ corresponds, as previously noted, to λ = 555 nm, and for twilight vision — to λ = 510 nm. At twilight vision, the eye can hardly distinguish colors, and radiation of different wavelengths is perceived as bluish grey. From which, maybe, there comes the proverb "All cats are gray in the dark",

The maximum of the curve of luminous efficacy of day vision corresponds to the maximum of solar radiation that has passed through the atmosphere and got to the Earth's surface (see § 27.4), which shows the expediency of the human eye structure.

A rod consists (fig. 29.21) of light-sensitive outer segment *1* and inner segment *2* containing the nucleus and mitochondria that provide for cellular functioning. Inside the outer segment, there are thin discs *3* with a diameter of about 6 μm. Each disc consists of a two-layer membrane and reminds a flattened liposome (see § 13.1). The visual discs have visual pigment, which is rhodopsin. In one cell, the number of disks is about several hundreds. The inner segment is connected to the nerve fiber.

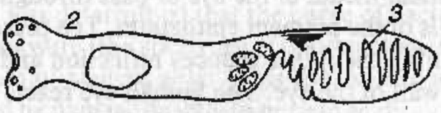


Fig. 29.21

Rhodopsin is a complex protein with a molecular mass of about 40 000. The diameter of its molecule is 4 nm if we assume its shape to be spherical.

Rhodopsin consists of opsin protein and chromophore group (retinal).

In general, retinal can have several spatial isomers, but only 11-cis-retinal binds opsin (fig. 29.22). Under the effect of light, retinal is separated from rhodopsin and transformed into a more stable conformation of a continuous trans-isomer.

Due to changes in the retinal structure, changes take place in the disc membrane; they are associated with changes in the position of rhodopsin. Rhodopsin moves from the interdiscal hydrophilic surface onto the inner hydrophobic phase of the membrane.

In darkness, if the disc membrane is impervious to Na^+, K^+, Ca^{2+} ions and others, then lighting conformational change in rhodopsin leads to a change

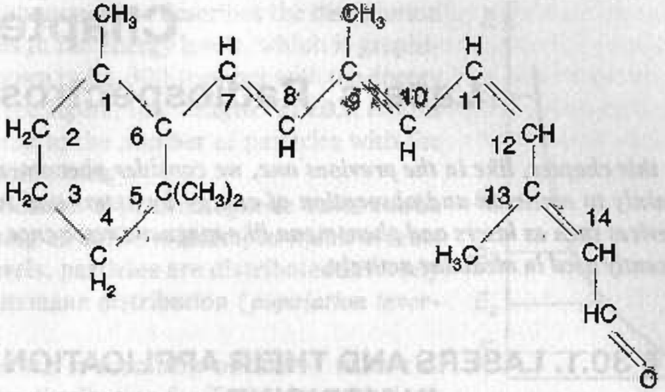


Fig. 29.22

in the state of the membrane: permeability for certain ions increases. In these processes, the function of rhodopsin is the following: under action of light, the material promotes formation of pores for some ions in the discs and closes the channels for sodium ions on the outer membrane. All this generates potentials that cause a nerve impulse. The peculiarity of the outer segments of the retinal rods is that in the dark, the potential has a sodium nature in contrast to the potential of other cells (see § 13.7). Because of changes in the structure of rhodopsin under the action of light, the permeability of membranes for sodium decreases dramatically, while it does not change in relation to other ions. In this case, permeability for potassium is of the most importance; the actual potential becomes potassic, and its polarity changes. It leads to the fact that in contrast to other known cells the potential is positive inside and negative outside on the cytoplasmic membrane of the outer segments of rods.

Similarly to rhodopsin, the pigment of cones contains 11-cis-retinal too, but the protein part of the pigment is different, so the pigments of cones are called iodopsins.

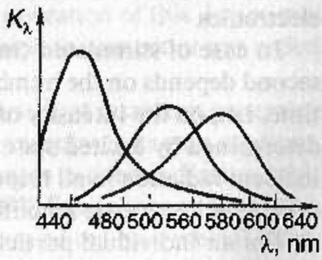


Fig. 29.23

Measurement of absorption spectra of certain cone varieties showed that each cone contains a definite type of iodopsin. Human cone iodopsins have the absorption maxima at 445, 535 and 570 nm (fig. 29.23). The three-component theory of color vision is based on these findings. In case of certain genetic diseases, synthesis of iodopsins is impaired, and eyes are not able to distinguish red and green colors (color blindness).

Chapter 30

Lasers. Radiospectroscop

In this chapter, like in the previous one, we consider phenomena relate mainly to emission and absorption of energy by atoms and molecul Devices such as lasers and phenomena like magnetic resonance have been recently used in medicine actively.

§ 30.1. LASERS AND THEIR APPLICATION IN MEDICINE

Despite the common nature of light and radio waves, for many yea optics and electronics developed independently. Light sources — excited particles and radio wave generators — seemed to have little in common. Only since the middle of the past century, there have been works about molecu amplifiers and generators of radio waves, which began the new independ field of physics — quantum electronics.

Quantum electronics **studies methods of amplification and generation of electro-magnetic oscillations using stimulated emission of quantum system** (see § 29.1). Achievements in this area of knowledge are more widely used in medicine.

Let us familiarize ourselves with certain phenomena underlying quantu electronics.

In case of stimulated (induced) emission, the number of transitions per second depends on the number of photons entering a substance over the same time, i.e., on the intensity of light. In addition, stimulated transitions will be determined by excited state population. Stimulated radiation is identical to incident radiation in all respects, including the phase, so we can say that electromagnetic wave are amplified coherently.

For an individual particle stimulated absorption, if the particle is in the ground state, and emission, if the particle is excited, are equally probable (see fig. 29.1, b). Therefore, even if in a material, the number of excited particles is equal to the number of unexcited particles, the incident electromagnetic wave will not be amplified.

In reality, in the ordinary state of a material, the number of unexcited particles is higher than that at any excited level. Thus, conditions for wave amp fication are even worse.

The Boltzmann law describes the distribution of particles in the energy levels, which is graphically, is shown in fig. 30.1 together with the energy levels. In the figure, the "length" of each level is proportional to the number of particles with the corresponding energy.

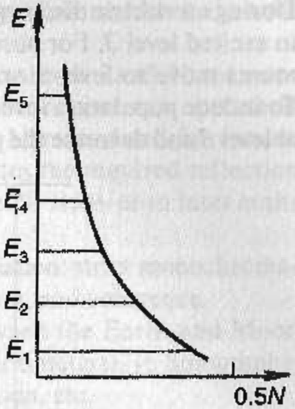


Fig. 30.1

Amplification of electromagnetic waves can be caused using an active medium, in which at least for two levels, particles are distributed inversely to the Boltzmann distribution (*population inversion*).

This state is formally obtained from the Boltzmann distribution for $T<0$ K, that is why it is called s state with a negative temperature. As light propagates in such a medium, its intensity increases, there is a phenomenon inverse to absorption. It means that in Bouguer's law (29.3) $k<0$, therefore, the population inversion corresponds to a medium with a negative absorption index.

A state with a population inversion can be induced by selecting the corresponding particles or exciting the particles intentionally, for example, by light or electric discharge. The state with a negative temperature does not exist for a long time by itself.

The phenomenon of stimulated radiation is used in quantum generators (amplifiers). In 1955, in the UHF range, the first such generator (maser) was independently designed by Soviet scientists N.G. Basov and A.M. Prokhorov and American scientists C. Townes et al.[1] Since operation of this device was based on stimulated emission of ammonia molecules, the generator was called ***molecular***.

The year 1960 saw the development of the first first quantum generator with visible radiation range. It was called ***laser***[2] with ruby as its gain medium. This optical quantum generator (OQG) generates pulsed radiation with a wavelength of 694.3 nm and pulse power of 1 MW. Excitation, or in the terminology of quantum electronics, pumping, is carried out by a special lamp.

In the same year, a gas helium-neon laser was created, in which, excitation was generated by an electric discharge. Neon atoms are emitting ones; helium atoms are auxiliary. Fig. 30.2 shows energy levels of helium and neon atoms.

---

[1] In 1964, N.G. Basov, A.M. Prokhorov, and C. Townes were awarded the Nobel Prize for their works.

[2] The name laser is an abbreviation of English words Light Amplification by Stimulated Emission of Radiation.

During an electric discharge, a part of neon atoms passes from t ground level *1* to excited level *3*. For pure neon, at this level, the lifetime is short, and the atoms move to level *1* or *2*, the Boltzmann distribution is implement To induce population inversion one should somehow increase the populati at level *3* and decrease the population at level *2*.

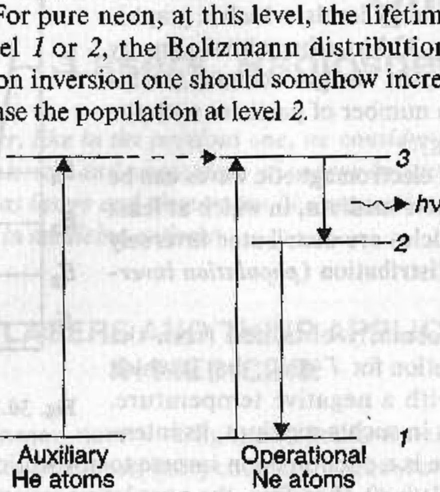


Fig. 30.2

Helium atoms contribute to increasing the population of level 3. The first excited level of helium coincides with level 3 of neon, therefore, whe an excited helium atom collides with an unexcited neon atom, energy transf takes place.

To unload level *2*, one selects the appropriate size for the gas-dischar tube, so that upon collision with its walls, a neon atom emits energy, movin from level *2* onto level *1*.

In such a way, stationary population inversion of neon levels *2* and *3* is ensured.

The principal structural element of the helium-neon laser (fig. 30.3) is dis- charge tube *1*, usually from quartz, with a diameter of about 7 mm. Tube *2*, at a pressure about 150 Pa, contains a mixture of helium and neon (the anount of helium is about 10 times greater than neon). In the tube, there are electrodes *3* to generate a gas discharge. At the ends of the tube, there are plane parall mirrors *4* and *5*, one (*5*) of them is semitransparent. Depending on the direction of their motion, pho- tons occurring upon stimulat emission either move out from the side surface of the tube, or cause stimulated transitions themselv being repeatedly reflected fro

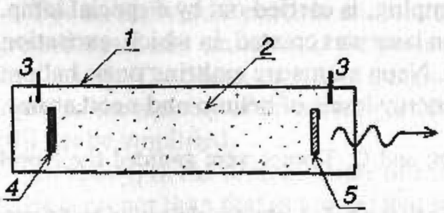


Fig. 30.3

the mirrors. Thus, the perpendicular beam will multiply and leave the tube through semitransparent mirror *5*.

This is a continuously operating laser. Since neon levels *2* and *3* have complex structure (not shown in fig. 30.2), helium-neon laser can operate at 30 wavelengths in the visible and infra-red range. Mirrors *4* and *5* are made with multilayer coatings, and interference generates the required reflection coefficient for the selected wavelength. Thus, a red helium-neon laser emits a wavelength of 632.8 nm.

Use of lasers is based on properties of their radiation: strict monochromaticity ($\Delta\lambda \approx 0.01$ nm), quite high power, narrow beam, and coherence.

Lasers are used to measure the distance between the Earth and Moon (the resulting error is in the order of dozens of centimeters), in holography, for burning small holes, as a means of communication, etc.

Laser are applied in medicine too. There are two main directions.

The first is based on the property of lasers to destroy biological tissues, which together with protein coagulation, makes it possible to perform certain bloodless dissections. In this sphere, the following applications can be noted: non-surgical treatment of retinal detachment (for this purpose, a special laser device — *ophthalmic coagulator* — is built); laser knife in surgery, which does not need sterilization; treatment of glaucoma by piercing openings of 50–100 μm for outflow of intraocular fluid; destruction of cancer cells; destruction of dentin in dentistry.

The second direction is related to holography (see § 24.8). For instance, ***gastroscopes*** based on a helium-neon laser with use of fiber optics were developed, permitting a three-dimensional holographic image of the stomach internal cavity.

Helium-neon laser is used to treat a number of diseases (trophic ulcers, coronary heart disease, etc.).

§ 30.2. SPLITTING OF ENERGY LEVELS OF ATOMS IN A MAGNETIC FIELD

In § 16.1 and 16.2, it was shown that a circuit with a current placed in a magnetic field is affected by a moment of force. When the circuit is in a stable equilibrium state, its magnetic moment coincides with the direction of the magnetic induction vector. A circuit left to its own resources and having a current gets into this state. The magnetic moments of the particles take quite different orientation in a magnetic field. Let us consider this problem from the quantum mechanics viewpoint.

In § 28.6, it was noted that a projection of the electron angular momen onto a certain direction takes discrete values. In order to detect these proje tions, it is necessary somehow to single out direction Z. One of the most common methods is to generate a magnetic field. In this case, the projection of th orbital angular momentum [see (28.26)], spin projection (28.27), projectio of the total angular momentum of the electron [see (28.30)] and projection of angular momentum of the atom [see (28.36)] onto the direction of magne induction vector B are determined.

The relationship between the angular momentum and magnetic moment (16.54) permits using the foregoing formulas to find discrete projectio of the corresponding magnetic moment onto the direction of vector B. Thu in contrast to classical ideas, the magnetic moments of particles are orient relative to the magnetic field angularly.

For an atom, for example, from (28.36), we obtain the following valu of projections of the magnetic moment on the direction of the magnetic induction vector:

$$p_{mz} = g\frac{e}{2m}L_{A_z} = g\frac{e}{2m}\frac{h}{2\pi}m_J = -g\mu_{Bo}m_J, \qquad (30.1)$$

where $\mu_{Bo} = eh/(4\pi m)$ is the ***Bohr magneton***; g is the ***Landé g-factor***, which for a given level of the energy of an atom, depends on quantum numbers L, J, S. In (30.1), the minus sign is due to the negative charge of the electron.

Formula (16.21) for the energy of a circuit with a current in a magne field can be applied to an atom. Taking into account that $p_m \cos\alpha$ is equal to projection p_{mz} of the magnetic moment onto the direction of the magne induction vector, we obtain:

$$E_p = -p_m B\cos\alpha = -p_{mz}B = g\cdot\mu_{Bo}\cdot m_J\cdot B.$$

Thus, in a magnetic field, the energy of an atom, taking into account that in its absence the energy of the atom is E_0, is determined by the formula:

$$E = E_0 + g\cdot\mu_{Bo}\cdot m_J\cdot B. \qquad (30.2)$$

Since the magnetic quantum number m_J ca take $2J + 1$ values from $+J$ to $-J$, then fro (30.2), it follows that when an atom is placed in a magnetic field, each energy level is split into $2J + 1$ sub-levels. This is sketchily shown in fig. 30.4 for $J = 1/2$. The energy spacing between adjacent sub-levels is:

$$\Delta E = g\cdot\mu_{Bo}\cdot B. \qquad (30.3)$$

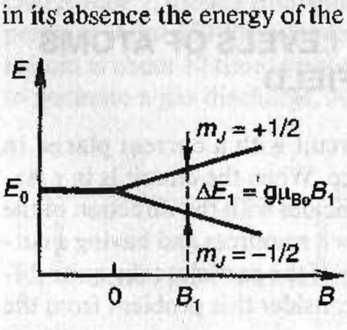


Fig. 30.4

Splitting of energy levels leads to splitting of spectral lines of atoms placed in a magnetic field. This phenomenon is called the *Zeeman effect.*

Let us write down expression (30.2) for two sublevels E_1 and E_2 that originated due to the presence of a magnetic field:

$$E_1 = E_{01} + g\mu_{Bo}Bm_{J1}, \quad E_2 = E_{02} + g\mu_{Bo}Bm_{J1}, \tag{30.4}$$

where E_{01} and E_{02} atomic energies in the absence of a magnetic field. Using (29.1) and (30.4), we obtain an expression for the frequencies emitted by the atom:

$$\nu = (E_{02} - E_{01})/h + \mu_{Bo}B(gm_{J2} - gm_{J1})/h = \nu_0 + \Delta\nu, \tag{30.5}$$

where

$$\nu_0 = (E_{02} - E_{01})/h \tag{30.6}$$

is the spectral line frequency in the absence of a magnetic field:

$$\Delta\nu = \mu_{Bo}B(g_2 m_{J2} - g_1 m_{J1})/h \tag{30.7}$$

is splitting of the spectral line in a magnetic field. From (30.7), it is clear that $\Delta\upsilon$ depends on the magnetic quantum number, Landé factor, and magnetic induction of the field. If $g_1 = g_2 = g_3$, then:

$$\Delta\nu = g\mu_{Bo}B(m_{J2} - m_{J1})/h.$$

According to rules of selecting the magnetic quantum number, we have:

$$\Delta m_J = m_{J2} - m_{J1} = 0, \pm 1. \tag{30.8}$$

It corresponds to the following three possible frequencies: $\nu_0 + g\mu_{Bo}B/h$, ν_0, $\nu_0 - g\mu_{Bo}B/h$, i.e., in a magnetic field, the spectral line is split into a triplet (fig. 30.5). Such splitting is called ***normal***, or ***simple Zeeman effect***, it is present only in powerful magnetic fields.

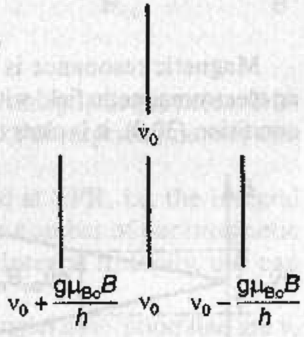

ν_0
$\nu_0 + \frac{g\mu_{Bo}B}{h}$
ν_0
$\nu_0 - \frac{g\mu_{Bo}B}{h}$

Fig. 30.5

A weak magnetic fields has ***anomalous Zeeman effect***; in this case, $g_1 \neq g_2$ and the splitting of the spectral lines is much more complicated.

§ 30.3. ELECTRON PARAMAGNETIC RESONANCE AND ITS BIOMEDICAL APPLICATION

In an atom placed in a magnetic field, spontaneous transitions between sublevels of the same level are unlikely. However, such transitions are induced under the effect of an external electromagnetic field. A necessary condition is coincidence of the frequency of the electromagnetic field with the frequency of the photon corresponding to the energy difference between the split sublevels. In this case, it is possible to see absorption of the energy of the electromagnetic field. This phenomenon is called ***magnetic resonance.***

Depending on the type of magnetic moment carrier particles, ***electron paramagnetic resonance*** (EPR) and ***nuclear magnetic resonance*** (NMR) are distinguished.

EPR occurs place in materials containing paramagnetic particles: molecules, atoms, ions, radicals with a magnetic moment caused by electrons. The emerging Zeeman effect is explained by splitting of electron levels (whence the name electron resonance). EPR is most common in particles with a purely spin magnetic moment (in foreign literature, this type of EPR is sometimes called ***electron spin resonance***).

EPR was discovered by E.K. Zavoysky in 1944. In the first experiment resonance absorption by ions in iron salts was observed. Zavoysky was able to study a number of regularities of this phenomenon.

From expressions (29.1) and (30.3) we obtain the following condition of resonance absorption of energy:

$$h\nu = g\mu_{\mathrm{Bo}}B_{\mathrm{res}}. \qquad (30.9)$$

Magnetic resonance is present if a constant field with induction B_{res} and an electromagnetic field with frequency ν affect a particle simultaneously. From condition (30.9), it is clear that it is possible to identify resonance absorption in two ways: either at a constant frequency varying the magnetic induction gradually or at a constant magnetic induction, varying the frequency gradually. Technically, the first variant is more convenient.

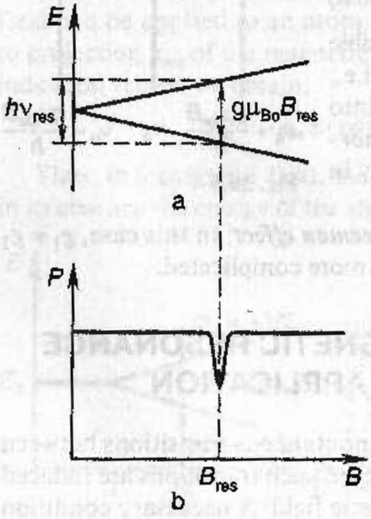


Fig. 30.6

Fig. 30.6 shows an electron energy level splitting (a) and electromagnetic wave power P change after passing through a sample depending on the value of magnetic field induction (b). Under condition (30.9), EPR occurs.

The shape and intensity of spectral lines observed at EPR are determined by interaction of the magnetic moments of electrons, in particular spin moments one with another, with the crystal grating of a solid, etc. Let us find out how these factors affect the nature of the spectra.

Let us assume that condition (30.9) is valid. For energy absorption, it is necessary that atoms of a material have higher population of the lower sublevels than the upper ones. Otherwise, stimulated energy emission will prevail.

In case of electron paramagnetic resonance, along with energy absorption and increase in higher sublevels population, the inverse process also takes place: non-radiative transitions to the lower sublevels; the energy of the particle is transferred to the crystal grating.

A process of energy transfer to the grating is called spin-grating relaxation; it is characterized by time τ. According to the Heisenberg relation (28.11), it leads to the level broadening.

Thus, resonance absorption does not occur exactly at one value of B, but within a certain interval ΔB (fig. 30.7). Instead of infinitely narrow absorption line, it will be a line of finite width: the shorter the spin-grating relaxation time, the greater the width of the line (in fig. 30.7, time $\tau_1 < \tau_2$ correspond to lines *1* and *2*).

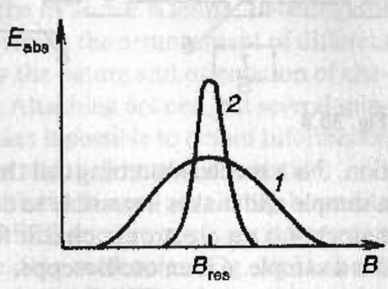


Fig. 30.7

Broadening of EPR lines depends also on electron spin interaction (spin-spin interaction) and on other interactiuns of paramagnetic particles. Different types of interaction affect not only the width of absorption line, but also its shape.

Under certain conditions, the energy absorbed at EPR, i.e. the integral (total) intensity of the line, is proportional to the number of paramagnetic particles. Whence it follows that by the measured integral intensity, one can judge about the concentration of these particles.

Important parameters characterizing singlet (single) absorption line are ν, B_{res}, g (location of the resonance point) corresponding to condition (30.9). At constant frequency ν, quantity B_{res} depends on g-factor. In the simplest case, a g-factor makes it possible to determine the nature of the system magnetism (spin or orbital). If an electron is bound to an atom, which is a part of a solid crystal lattice or molecular system, then strong internal fields will influence it. Measuring g-factor, it is possible to obtain information of the fields and intra-molecular bond.

However, if only a singlet absorption line was obtained in the experiment, many applications of magnetic resonance methods would be impossible. Most applications, including biomedical ones, are based on analysing a group of lines. In the EPR spectrum, the presence of a group of close lines is conventionally called *splitting*. There are two characteristic types of EPR spectrum splitting.

The first type — *electron splitting* — occurs when a molecule or atom has not one, but several electrons causing EPR. The second one — *hyperfine splitting* — is present upon interaction of electrons with the nuclear magnetic moment.

Modern methods of measuring EPR are based on determining the change in any parameter of the oscillatory system occurring upon absorption of electromagnetic energy.

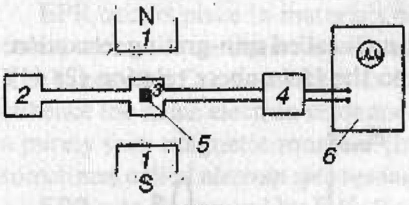


Fig. 30.8

An instrument used for this purpose is called EPR *spectrometer*. It consists of the following principal parts (fig. 30.8). 1 is an electromagnet that generates a strong homogeneous magnetic field, with gradually changing induction. *2* is a generator of UHF radiation. *3* is a special absorbing cell that focuses the incident UHF radiation onto a sample and makes it possible to detect its absorption of energy (volume resonator). *4* is an electronic circuit for observation or recording EPR spectra. *5* is a sample. *6* is an oscilloscope.

A modern home EPR Rubin spectrometer (fig. 30.9) uses the frequency of about 10 GHz (the wavelength is 0.03 m). It means, according to (30.9), that for $g = 2$, the maximum EPR absorption is present at $B = 0.3$ T.

In practice, EPR spectrometers record not energy absorption curve (fig. 30.10, a), but its derivative (fig. 30.10, b).

One of biomedical applications of the EPR method is detection and research of free radicals. For example, EPR spectra of irradiated proteins permit

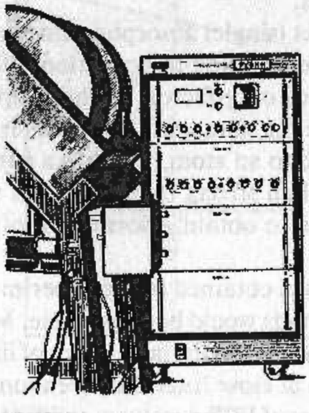


Fig. 30.9

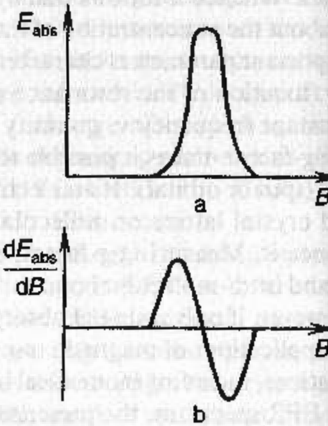


Fig. 30.10

explaining the mechanism of free radical formation and, in this connection, monitoring changes in primary and secondary products of radiation injury.

EPR is widely used to study photochemical processes, in particular, photosynthesis. Carcinogenic activity of certain materials is studied too. For sanitary purposes, the EPR method is used to determine the concentration of radicals in the air.

A method of spin labels was proposed relatively recently, specifically for study of biological molecules; its essence is that a paramagnetic material with a well-known structure is bound to a molecule of the object under study. By EPR spectra, the position of such a spin label in the molecule is found. Labeling different molecular parts, it is possible to determine the arrangement of different groups of atoms, their interactions, to study the nature and orientation of chemical bonds and to detect molecular motion. Attaching not one, but several spin-labels to the molecule, for example, two, makes it possible to obtain information about the distance between the labeled groups and their mutual orientation.

Spin probes — paramagnetic particles that are non-covalently bound to molecules — are also used. Changes in the EPR spectrum of spin probes give information about the state of the surrounding molecules. Fig. 30.11 shows EPR spectra of nitroxyl radical, which is placed in glycerol as a spin probe. As the temperature increases, the viscosity of glycerol decreases, and this changes the shape of the EPR spectrum. Thus, the shape of the EPR spectrum can be used to determine microviscosity, which is the viscosity of the spin probe's direct medium. In this way, in particular, it is possible to determine the microviscosity of membrane lipid layer (see § 13.2).

In Russia, biological objects are widely researched by the EPR method.

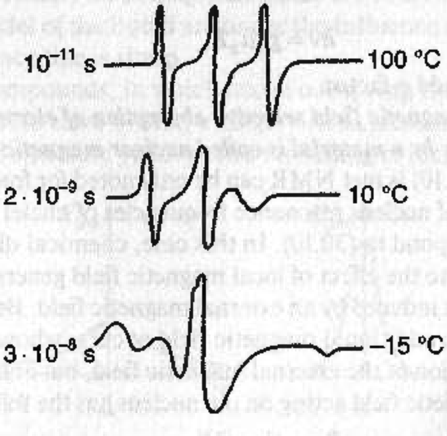


Fig. 30.11

§ 30.4. NUCLEAR MAGNETIC RESONANCE (NMR). NMR-INTROSCOPY

Nuclear magnetic resonance does not belong to the section of physics of atoms and molecules, but is considered in chapter dealing with EPR as a phenomenon of magnetic resonance.

Nuclear magnetic moment is a sum of the magnetic moments of nucleons. Usually, this moment is expressed in nuclear magnetons (μ_n); 1 μ_n = $= 5.05 \cdot 10^{-27}$ A $\cdot$ m^2. The magnetic moment of proton is approximately equal to $p_{mp} = 2.79\ \mu_n$, the magnetic moment of neuron is approximately equal to $p_{mn} = -1.91\ \mu_n$. The negative sign means that the neutron's or nuclear magnetic moment is oriented opposite to their spin.

Let us represent magnetic moments of some nuclei (Table 30.1).

Table 30.1

	${}^4_2$He	${}^9_4$Be	${}^{12}_6$C	${}^{115}_{49}$In
μ_n	0	−1.2	0	5.5

The magnetic moment of a nucleus placed in a magnetic field can take only a discrete orientation. This means that the energy of the nucleus will be corresponded by the sublevels, the spacing between which depends on the induction of the magnetic field.

If under these conditions the nucleus is affected by electromagnetic field, transitions between sublevels can be induced. To accomplish these transitions as well as absorption of electromagnetic energy, it is necessary to fulfill a condition like (30.9):

$$h\nu = g_n \mu_n B, \tag{30.10}$$

where g_n is the Landé g-factor.

In a constant magnetic field selective absorption of electromagnetic waves of a definite frequency by a material is called nuclear magnetic resonance **(NMR).**

If condition (30.10) is met NMR can be only noted for free atomic nuclei. Experimental values of nucleus resonance frequencies of nuclei of atoms and molecules do not correspond to (30.10). In this case, chemical displacement occurs, which happens fue to the effect of local magnetic field generated inside the atom by electron currents induced by an external magnetic field. Because of such a diamagnetic effect, an additional magnetic field occurs, whose induction proportional to the induction of the external magnetic field, but antiparallel to it, so the total effective magnetic field acting on the nucleus has the following induction:

$$B_{ef} = (1-\sigma)B, \tag{30.11}$$

where σ is shielding constant, with an order of magnitude of 10^{-6} and depending on the electronic environment of the nucleus.

Whence it follows that for this type of nuclei located in different environments (different molecules or different, but not equivalent places in the same molecule), the resonance is present at different frequencies. This exactly determines chemical displacement. It depends on the nature of the chemical bond, electronic structure of the molecules, concentration of the material, type of solvent, temperature, etc.

In a molecule, if several nuclei are shielded differently, i.e., the nuclei occupy chemically nonequivalent positions, they have a different chemical displacement. NMR spectrum of such a molecule contains as many resonance lines as many chemically nonequivalent groups of nuclei of this type are in the molecule. The intensity of each line is proportional to the number of nuclei in the group.

In NMR spectra, two types of lines are distinguished by their width. Spectra of solids have a large width, and this sphere of NMR application is called *wide-line* NMR. Narrow lines are present in liquids, and this phenomenon is called *high-resolution* NMR.

Fig. 30.12 shows nuclear magnetic resonance curves for solids (a) and liquids (b). In liquids, the sharp peak is due to the following. Each nucleus interacts with its neighbors. Since the orientation of the nuclear magnetic moments surrounding a nucleus of this type varies from one point to another in the material, the total magnetic field acting on different nuclei of the same type also changes. This means that for the total population of nuclei, the resonance region should be a wide line. However, in a liquid, due to the quick movements of molecules, local magnetic fields are short-lived. This leads to the fact that the nuclei of the liquid are under the influence of the same average field, so the resonance line is sharp.

For chemical compounds, in which nuclei occupying chemically equivalent places in the molecule show NMR, a single line is present. Compounds with a more complicated structure yield spectra consisting of many lines.

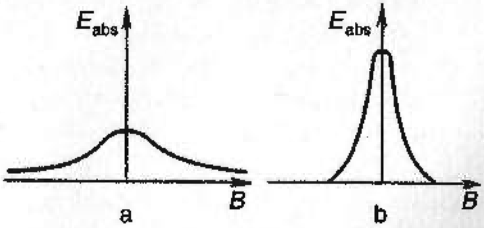


Fig. 30.12

The structure of molecules can be determined by chemical displacement the number and position of spectral lines.

Chemists and biochemists widely use the NMR method to study the structure of materials, beginning from the simplest molecules of inorganic materials to the most complex molecules of living objects, as well as in solving many problems connected with the course of chemical reactions, study of parent materials structures, and final products of reactions. An advantage of this analysis is that it does not destroy the objects of study the way it happens, for example, in chemical analysis.

At many points of a sample, determination of NMR spectrum parameters can offer very interesting possibilities for medicine. One can obtain the complete picture of spatial distribution of molecules containing, for example, hydrogen or phosphorus atoms (for protons or phosphorus nuclei magnetic resonance respectively) passing gradually layer by

layer through the entire sample (scanning).

All this is carried out without destroying the sample, and, therefore, it is possible to research living objects. This method is called NMR-introscopy (see § 24.8), it permits differentiating bones, blood vessels, normal tissues and tissues with malignization. NMR-introscopy makes it possible to distinguish images of soft tissues; for example, it distinguishes the image of grey and white cerebral matter, tumor cells from healthy ones; while minimum sizes of pathological inclusions can be a fraction of millimeter. NMR-introscopy is expected to become an effective method to diagnose diseases caused by abnormality of organs and tissues.

The frequency of electromagnetic waves causing transitions between energy states in EPR and NMR corresponds to the radio frequency band. Therefore, both these phenomena are related to radiospectroscopy.

Section 8

IONIZING RADIATION. FUNDAMENTALS OF DOSIMETRY

Ionizing radiation is flows of particles and electromagnetic quanta, whose interaction with a medium leads to ionizing of its atoms and molecules. Ionizing radiations are X-ray and γ-radiation, flows of α-particles, electrons, positrons, protons, neutrons. This section describes both ionizing radiation sources (X-ray tubes, radionuclides[1], particle accelerators) and physical aspects of applying this radiation for medical purposes. Physicians and biologists are undoubtedly interested in such problems as interaction of ionizing radiation with materials and elements of ionizing radiation dosimetry.

---

[1] *Nuclides* are atoms whose nuclei differ in their composition, i.e. contain different numbers of protons or (and) neutrons. In some cases, this term is used as a common name of atomic nuclei. *Radionuclides* are nuclides capable of radioactive decay.

Chapter 31

X-ray (roentgen) radiation

X-radiation is electromagnetic waves with a wavelength of approximately from 80 to 10^{-5} nm. X-radiation with the longest wavelength is overlapped by short-wave ultraviolet radiation, the short wavelength X-radiation is overlapped by long wavelength γ-radiation. According to the method of excitation, X-radiation is subdivided into continuous and characteristic.

§ 31.1. X-RAY TUBE STRUCTURE. BREMSSTRAHLUNG X-RAYS

The most common source of X-ray is an X-ray tube, which is a two-electrode vacuum device (fig. 31.1). Heating cathode *1* emits electrons *4*. Anode *2*, often called anti-cathode, has an inclined surface in order to direct resulting X-ray radiation *3* at an angle to the horizontal axis of the tube. The anode is made from a good conductive material to remove the heat generated by electron impact. The anode surface is made from refractory materials with a large atomic number in the Mendeleev' periodic table, for example, tungsten. In certain cases, the anode is intentionally cooled with water or oil.

For diagnostic tubes, it is important that the X-ray source be point; this can be achieved by focusing electrons in the same anti-cathode place. Thus, two opposite tasks should be structurally taken in account: on the one hand, the electrons should strike the same place of the anode; on the other hand, in order to prevent overheating, it is desirable to distribute the electrons across different areas of the anode. An interesting technical solution is an X-ray tube with a rotating anode (fig. 31.2).

Due to deceleration of an electron (or other charged particle) by the electrostatic field of the atomic nucleus and atomic electrons of the anti-cathode material, *bremsstrahlung (deceleration) radiation occurs.*

Its mechanism can be explained in the following way. The magnetic field, whose induction depends on the velocity of an electron, is connected with a moving electric charge. Upon deceleration, magnetic induction decreases and, according to Maxwell's theory, an electromagnetic wave occurs.

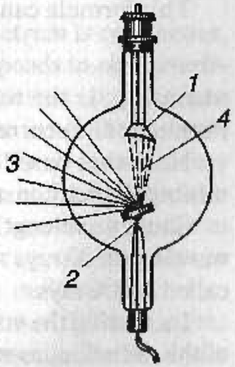


Fig. 31.1

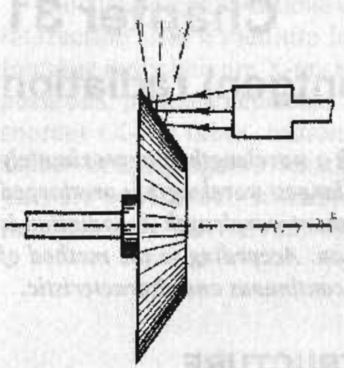


Fig. 31.2

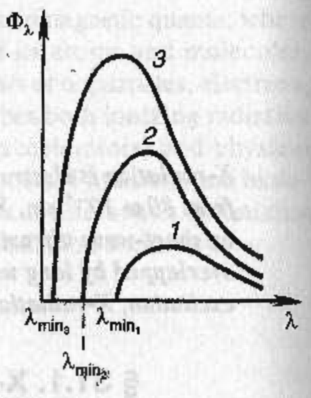


Fig. 31.3

When the electron decelerates, only a part of the energy is transformed to a photon of X-ray radiation, the other part heats the anode. Since the relationship between these parts is a random value, a large number of electrons decelerate and form a continuous X-ray spectrum. That is why bremsstrahlung radiation is also described as continuous radiation. Fig. 31.3 shows dependences of X-ray flow on wavelength λ (spectra) at different voltages at the X-ray tube: $U_1 < U_2 < U_3$.

In each spectrum, continuous radiation with the shortest wavelength λ_{min} occurs when the energy taken by an electron in the accelerating field is completely converted into a photon energy:

$$eU = h\nu_{max} = hc/\lambda_{min}, \tag{31.1}$$

whence

$$\lambda_{min} = hc/(eU). \tag{31.2}$$

This formula can be transformed into a more practical expression:

$$\lambda_{min} = 1.23/U, \tag{31.3}$$

where λ_{min} is the wavelength as a multiple of 10^{-10} nm; U is voltage in kV. Formula (31.3) corresponds to fig. 31.3.

Note that based on (31.2), one of the most accurate methods of determining Planck constant experimental is developed.

Short wavelength X-rays have usually a greater permeability than long wavelength X-rays and are called *hard* X-rays; long wavelength X-rays are called *soft* X-rays.

Increasing the voltage at the X-ray tube, one changes spectral composition of the radiation, as we can see from fig. 31.3 and formula (31.3), and the X-ray hardness is increased.

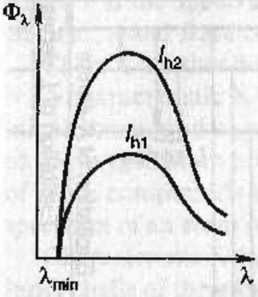


Fig. 31.4

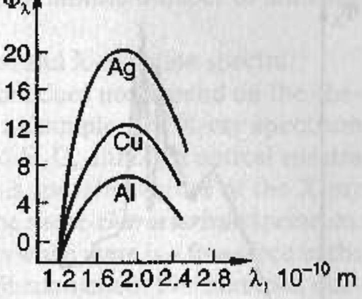


Fig. 31.5

If we increase the temperature of the cathode, in the tube, the emission of electrons and current intensity will increase too. This will lead to an increase in the number of photons of X-ray per second. Its spectral composition will remain the same. Fig. 31.4 shows the spectra of bremsstrahlung radiation at the same voltage but at different cathode heating current intensity; $I_{h1} < I_{h2}$.

The X-ray radiant flux is calculated by the following formula:

$$\Phi = kIU^2Z, \tag{31.4}$$

where U and I are respectively voltage and current intensity in the X-ray tube; Z is the atomic number for the anode material; k is the proportionality coefficient. Spectra obtained from different anti-cathodes at equal U and I_h are shown in fig. 31.5.

§ 31.2. CHARACTERISTIC X-RADIATION. ATOMIC X-RAY SPECTRA

Increasing the voltage at the X-ray tube, while the spectrum is continuous, one can notice occurrence of a line spectrum, which corresponds to *characteristic X-radiation* (fig. 31.6). It occurs due to accelerated electrons passing into the depths of an atom and knocking out electrons from the inner shells. From the upper levels, electrons move to free positions (fig. 31.7); as a result, they generate photons of characteristic radiation. As we can see in the figure, characteristic X-radiation consists of series K, L, M etc, the name of which serves to denote the electronic shells. Since upon emission K-series vacate spaces in the higher shells, the lines of the other series are emitted simultaneously.

In contrast to optical spectra, characteristic X-ray spectra of different atoms are of the same type. Fig. 31.8 shows spectra of different elements. The similarity of these spectra types is due to the fact that in different atoms, the inner shells

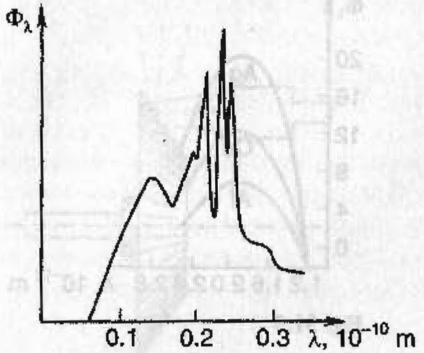


Fig. 31.6

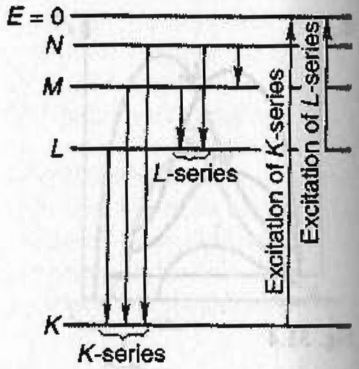

E = 0
N
M
L
K
L-series
K-series
Excitation of K-series
Excitation of L-series

Fig. 31.7

are the same and differ only in energy, since the force of the nucleus increases in proportion to increasing atomic number. This leads to the fact that characteristic spectra move towards a higher frequencies region as the nucleus charge increases. This regularity is seen in fig. 31.8, and is known as ***Moseley's law***:

$$\sqrt{\nu} = A(Z - B), \qquad (31.5)$$

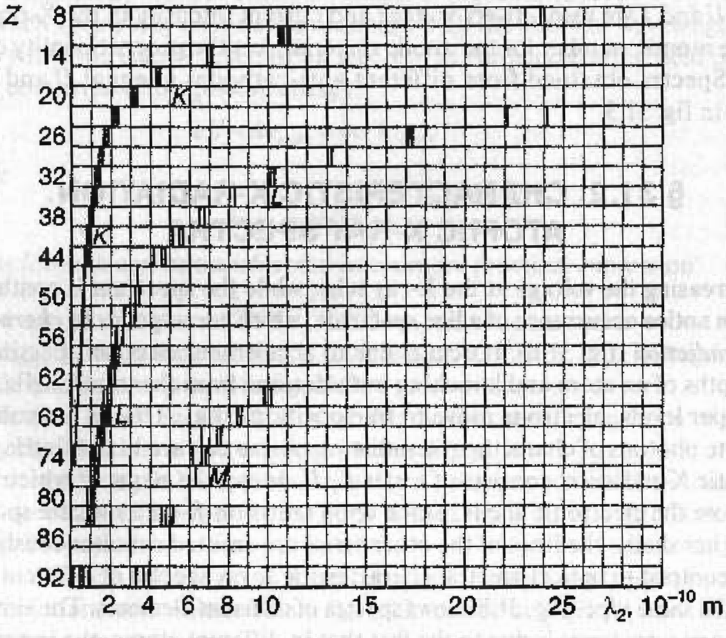


Fig. 31.8

where ν is the spectral line frequency; Z is the atomic number of emitting element; A and B are constants.

There is another difference between optical and X-radiation spectra.

A characteristic X-ray spectrum of an atom does not depend on the chemical compound comprising the atom. For example, the X-ray spectrum of the oxygen atom is the same for O, O_2 and H_2O, although optical spectra of these compounds are quite different. This specific feature of the X-ray spectrum of an atom served as the basis for the name ***characteristic*** spectrum.

Characteristic X-raradiation always occurs when there is a free place in the inner shells of the atom, irrespective of the phenomenon. For example, characteristic X-radiation accompanies a certain type of radioactive decay (see § 32.1), which consists in nucleus capturing the electron from the inner shell.

§ 31.3. INTERACTION OF X-RADIATION WITH MATERIALS

Primary interaction of the X-ray photon with electrons of atoms and molecules of a material determine the recording and use of X-ray as well as its impact on biological objects.

Depending on the ratio between photon energy $h\nu$ and ionizing energy[1] A_i, there are three main processes.

Coherent (classical) scattering

Scattering of long wavelength X-rays takes place mainly without a change in the wavelength; this is called *coherent* scattering. It occurs if the photon energy is less than ionizing energy: $h\nu < A_i$.

Since in this case the X-ray of photon energy and atom energy does not change, coherent scattering proper does not cause a biological effect. Nevertheless, when designing means of protection against X-ray, one should take into account that the direction of the primary beam can change. This type of interaction is important in X-ray crystallography (see § 24.7).

Incoherent scattering (the Compton effect)

In 1922, A.H. Compton observing the scattering of hard X-rays found a decrease in permeability of a scattered beam in comparison with the incident beam. This meant that the wavelength of scattered X-rays is higher than

---

[1] In this, ionizing energy is understood to be the energy required to remove internal electrons from an atom or molecule.

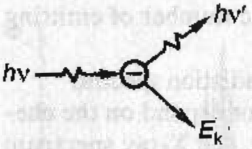


Fig. 31.9

that of incident X-rays. Scattering of X-ray with a change in its wavelength is called *incoherent* and the phenomenon itself is called the *Compton effect*. It occurs if the energy of X-ray photon is higher than the ionizing energy: $h\nu > A_i$:

This phenomenon is due to the fact that upon interaction with an atom, energy $h\nu$ of a photon is spent for formation of a new scattered photon of X-ray with energy $h\nu'$, separating the electron from its atom (ionizing energy A_i) and imparting kinetic energy E_k to the electron:

$$h\nu = h\nu' + A_i + E_k. \tag{31.6}$$

Since in many cases $h\nu \gg A_i$ and the Compton effect occurs in free electrons, it can be written approximately as:

$$h\nu = h\nu' + E_k. \tag{31.7}$$

It is essential that in this phenomenon (fig. 31.9) along with seconda X-rays (energy $h\nu'$ of the photon) there are recoil electrons (kinetic energy E_k of the electron). Thus, atoms or molecules become ions.

Photoelectric effect

In the photoelectric effect, an atom absorbs X-radiation, loses an electron and is ionized (photoionization).

Three principal interaction process discussed above are primary, they lead to subsequent secondary, tertiary etc. phenomena. For example, ionized atoms can emit a characteristic spectrum, excited atoms can become sources of visible light (X-ray luminescence), etc.

Fig. 31.10 shows a diagram of possible processes occurring because of X-ray penetrating a material. Several dozens of processes like the depicted one can occur before the energy of X-ray photon is converted into energy of molecular thermal motion. As a result, molecular composition of the material will change.

The processes represented by fig. 31.10 diagram underlie the phenome observed under the impact of X-rays on material. Here are some of them.

X-ray luminescence is glowing of a number of materials upon X-irradiatio This glow of platinocyanic barium permitted Roentgen to discover X-rays. This phenomenon is used to design special luminous screens for visual observation of X-rays as well to amplify the effect of X-rays on a photo plate.

Chemical action of X-rays, such as formation of hydrogen peroxide in water is well known. A practically important example is its effect on a photographic plate which makes it possible to record such rays.

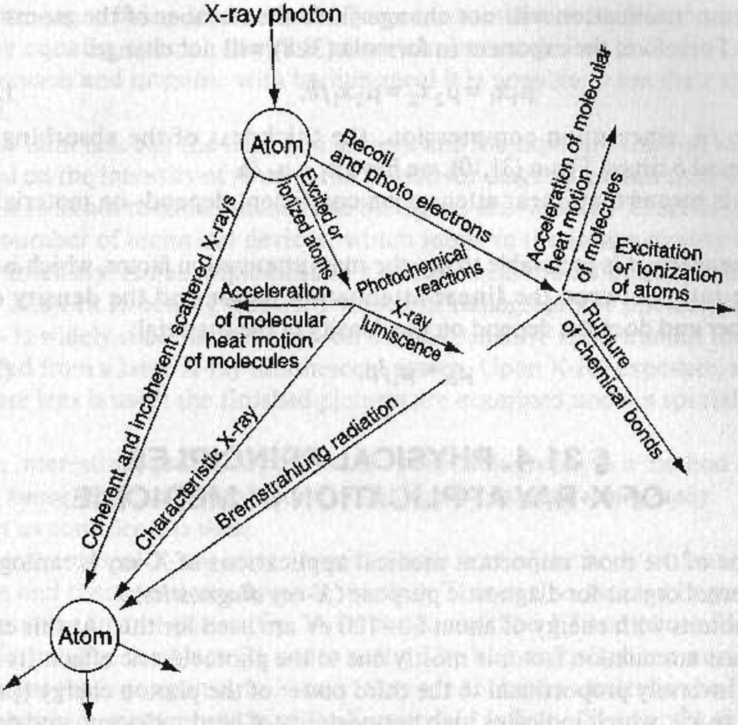

X-ray photon
Atom
Recoil and photo electrons
Excited or ionized atoms
Photochemical reactions
Acceleration of molecular heat motion of molecules
Excitation or ionization of atoms
Rupture of chemical bonds
Acceleration of molecular heat motion of molecules
X-ray lumiscence
Coherent and incoherent scattered X-rays
Characteristic X-ray
Bremstrahlung radiation
Atom

Fig. 31.10

The ionizing effect manifests as an increase in electrical conductivity under influence of X-rays. This property is used in dosimetry to quantify the effect of this type of radiation.

Because of many processes, the primary X-ray beam is attenuated in accordance with the law (29.3).

Let us write it in the following form:

$$I = I_0 e^{-\mu x}, \qquad (31.8)$$

where μ is *linear attenuation factor.* It can be represented as consisting of three summands corresponding to coherent scattering μ_c, incoherent scattering μ_{ic}, and photoelectric effect μ_{ph}:

$$\mu = \mu_c + \mu_{ic} + \mu_{ph}. \qquad (31.9)$$

The intensity of X-rays is attenuated proportionally to the number of atoms of the material, through which this beam passes. If we compress the material along axis *X*, for instance, *b* times, increasing its density *b* times too, then the beam attenuation will not change since the number of the atoms is the same. Therefore, the exponent in formula (31.8) will not change:

$$\mu_1 x_1 = \mu_2 x_2 = \mu_2 x_1 / b; \tag{31.10}$$

$x_2 = x_1/b$, since upon compression, the thickness of the absorbing layer decreased b times. From (31.10), we have $\mu_1 = \mu_2/b$.

This means that linear attenuation coefficient depends on material density.

Therefore, it is preferable to use the mass attenuation factor, which is equal to the ratio between the linear attenuation factor and the density of the absorber and does not depend on the density of the material:

$$\mu_m = \mu_1/\rho. \tag{31.11}$$

§ 31.4. PHYSICAL PRINCIPLES OF X-RAY APPLICATION IN MEDICINE

One of the most important medical applications of X-ray is radiography of internal organs for diagnostic purpose (***X-ray diagnostics***).

Photons with energy of about 60–120 eV are used for this. At this energy, the mass attenuation factor is mostly due to the photoelectric effect. Its quantity is inversely proportional to the third power of the photon energy (proportional to λ^3), which indicates high permeability of hard radiation, and proportional to the third power of the atomic number of the absorbing material:

$$\mu_m = k\lambda^3 Z^3, \tag{31.12}$$

where k is proportionality coefficient.

Absorption of X-rays scarcely depends on the compound in which the atom is represented in the material, so it is easy to compare mass attenuation coefficients μ_{mb} of bone $Ca_3(PO_4)_2$ and μ_{mw} of soft tissue or water H_2O using formula (31.12). The atomic numbers of Ca, P, O and H are respectively equal to 20, 15, 8 and 1. Substituting these numbers into (31.12), we obtain:

$$\frac{\mu_{mb}}{\mu_{mw}} = \frac{3\cdot 20^3 + 2\cdot 15^3 + 8\cdot 8^3}{2\cdot 1^3 + 8^3} = 68.$$

The sizeable difference in the absorption of X-ray by different tissues makes it possible to see images of inner organs in the shadow projection.

X-ray diagnostics is used in the two following variants. In ***fluoroscopy***, an image is examined on X-ray luminescent screen; in ***radiography***, an image is recorded on a photo film.

If the organ being studied and adjacent tissues attenuate X-rays approximately equally, special contrast materials are used. For example, filling the stomach and intestine with barium meal it is possible to see their shadow image.

The luminance of the image on ascreen and the exposure time on the film depend on the intensity of X-rays. If it is used for diagnostics, intensity cannot be high so as not to cause undesirable biological afte-reffects. Therefore, there are a number of technical devices, which improve the image quality at low X-ray intensity. An example of such a device is electronic optical converter (see § 27.8). In screening studies, a variant of radiography — photofluorography — is widely used; in this case, on a small sensitive film, a small image is recorded from a large X-ray luminescent screen. Upon X-ray exposure, a large aperture lens is used; the finished pictures are examined under a special magnifier.

An interesting, promising option of X-ray radiography is a method called ***X-ray tomography*** and its computerized variant, ***computer tomography***.

Let us consider this issue.

A conventional X-ray image covers a large area of the body where various organs and tissues overshadow one another. This can be avoided by periodically moving in anti-phase (fig. 31.11) the X-ray tube *XT* and photo film *Phf* together relative to object *Ob* under study. The human body has a number of inclusions that are opaque to X-rays; they are shown in circles in the figure. As we can see, at any position of the X-ray tube (*1*, *2* etc.) X-rays pass through the same point of the object, which is the center, with respect to which *XT* and *Phf* perform periodic motions. This point, a small opaque inclusion specifically, is shown by a dark circle. Its shadow image moves together with *Phf*, taking positions *1*, *2* etc. one after another. Other inclusions in the body (bones, indurations, etc.) create a certain general background on *Phf*, since X-rays are not overshadowed by them all the time. Changing the position of the center of oscillation, we can obtain a layered X-ray image of the body. From where, the name of ***tomography*** (layered recording) originated.

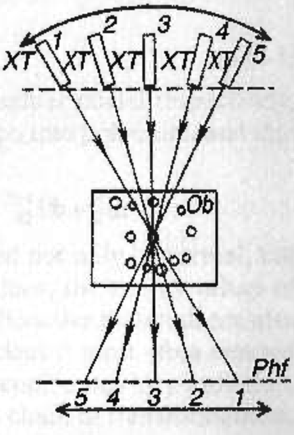


Fig. 31.11

It is possible to process a shadow X-ray image during tomography using a narrow X-ray beam, a screen (instead of *Phf*) consisting of semiconductor detectors of ionizing radiation (see § 32.5) and a computer. This modern version of tomography (computational or computer X-ray tomography) makes it possible to obtain layered images of the body on the screen of an electron beam tube or on paper with details, which are less than 2 mm with a margin of X-ray absorption of down to 0.1%. This makes it possible, for example, to distinguish gray and white cerebral matter and to see very small tumor masses.

The first Nobel Prize was awarded to W. Roentgen (1901); and in 1979 the Nobel Prize was awarded to G. Hounsfield and A. Cormack for development of computer X-ray CT-scanner.

For therapeutic purposes, X-rays are used mainly for destruction of malignant neoplasms (***radiotherapy***).

Chapter 32

Radioactivity. Interaction of ionizing radiation with materials

One of the most common causes of ionizing radiation is radioactive decay of atomic nuclei. This chapter also deals with interaction of ionizing radiation with materials.

§ 32.1. RADIOACTIVITY

***Radioactivity is spontaneous decay of unstable nuclei accompanying emission of other nuclei or elementary particles.* A characteristic feature that distinguishes it from other types of nuclear transformations is spontaneity of this process.** There is natural and induced radioactivity.

Natural radioactivity is present in unstable naturally occurring nuclei. *Induced radioactivity* is the radioactivity of nuclei farising due to different nuclear reactions. There is no fundamental difference between natural and induced radioactivity. They have common regularities.

Let us consider the main types of radioactive decay.

α-Decay *consists in spontaneous transformation of* α*-nucleus accompanying an* alpha *particle emission*. The formula of α-decay, taking into account the displacement rule, is written as:

$$^{A}_{Z}\mathrm{X} \rightarrow {}^{A-4}_{Z-2}\mathrm{Y} + {}^{4}_{2}\alpha, \tag{32.1}$$

where X and Y are symbols of the parent and daughter nuclei respectively. An example of alpha decay is transformation of radon into polonium, and that of polonium into lead:

$$^{222}_{86}\mathrm{Rn} \rightarrow {}^{218}_{84}\mathrm{Po} + {}^{4}_{2}\alpha \quad \text{and} \quad {}^{218}_{84}\mathrm{Po} \rightarrow {}^{214}_{82}\mathrm{Pb} + {}^{4}_{2}\alpha.$$

At α-decay, the daughter nucleus can be formed not only in normal, but also in excited states. Since they take discrete values, the energy values of α-particles, leaving different nuclei of the same radioactive material, are also discrete. The excitation energy of the daughter nucleus is most often emitted in the form of γ-photons. That is why α-decay is accompanied by γ-radiation.

If the daughter nuclei are radioactive, there is a chain of transformations, the final link of which is a stable nucleus.

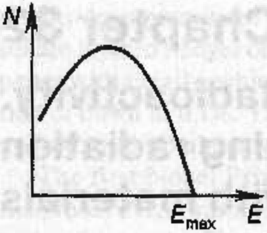


Fig. 32.1

β-*Decay* consists in intra-nuclear mutual transformation of a neutron and proton. There are three types of beta decay.

1. Electron, or β^--decay: β^--particle (electron) leaving the nucleus. Energies of beta particles take every possible values from 0 to E_{max}, the energy spectrum is continuous (fig. 32.1). This phenomenon does not correspond to discrete nuclear energy states. In 1932, W. Pauli suggested that at the time, together with a β^--particle, another neutral particle with a very small mass leaves the nucleus. According to E. Fermi's idea, this particle was called neutrino. Later it was established that neutrino occurs at β^+-decay, and at β^--decay, anti-neutrino occurs.

The energy emitted at beta decay is distributed between a β-particle and neutrino (or anti-neutrino).

The formula of β^--decay, taking into account the displacement rule, is:

$${}^{A}_{Z}\mathrm{X} \rightarrow {}^{A}_{Z+1}\mathrm{Y} + {}^{0}_{-1}\beta + \tilde{\nu}, \qquad (32.2)$$

where ν is the designation of anti-neutrino. An example of a β^--decay is transformation of tritium into helium:

$${}^{3}_{1}\mathrm{H} \rightarrow {}^{3}_{2}\mathrm{He} + {}^{0}_{-1}\beta + \tilde{\nu}.$$

At β^--decay, an electron is formed due to intra-nuclear transformation of a neutron into a proton:

$${}^{1}_{0}n \rightarrow {}^{1}_{+1}p + {}^{0}_{-1}\beta + \tilde{\nu}. \qquad (32.3)$$

2. Positron, or β^+ -decay has the following formula:

$${}^{A}_{Z}\mathrm{X} \rightarrow {}^{A}_{Z-1}\mathrm{Y} + {}^{0}_{+1}\beta + \nu, \qquad (32.4)$$

where ν is the designation of neutrino. An example of β^+-decay is transformation of rubidium into krypton:

$${}^{79}_{37}\mathrm{Rb} \rightarrow {}^{79}_{36}\mathrm{Kr} + {}^{0}_{+1}\beta + \nu,$$

At β^+-decay, a positron occurs due to intra-nuclear transformation of a proton into a neutron:

$${}^{1}_{+1}p \rightarrow {}^{1}_{0}n + {}^{0}_{1}\beta + \nu. \qquad (32.5)$$

3. Electron capture, or *e*-capture: this type of radioactivity consists in capture of inner electrons of an atom by the nucleus, because of which a proton of the nucleus transforms into a neutron:

$${}^{1}_{+1}p + {}^{0}_{-1}\beta \rightarrow {}^{1}_{0}n + \nu. \qquad (32.6)$$

The formula of electron capture:

$$ {}_{Z}^{A}X + {}_{-1}^{0}\beta \rightarrow {}_{Z-1}^{A}Y + \nu. \tag{32.7}$$

An example of *e*-capture can be transformation of beryllium into lithium:

$$ {}_{4}^{7}Be + {}_{-1}^{0}\beta \rightarrow {}_{3}^{7}Li + \nu. $$

Depending on the inner shell, from which an electron is captured, sometimes they distinguish *K*-capture, *L*-capture, etc. At electron capture, free spaces in the electron shell are formed, so this type of radioactivity is accompanied by characteristic X-radiation. It is in X-radiation studies that electron capture was established.

Radioactivity is also spontaneous nuclear fission, proton radioactivity, etc. The concept of radioactivity is sometimes extended to transformations of elementary particles.

§ 32.2. BASIC LAW OF RADIOACTIVE DECAY. ACTIVITY

***Radioactive decay* is a statistic phenomenon.** It is impossible to foresee when a given unstable nucleus will disintegrate, one can only make certain probabilistic judgments about this event. For a large population of radioactive nuclei, a statistical law expressing dependence of the amount of undecayed nuclei on time can be obtained.

Let dN *nuclei* decay over a sufficiently small time interval dt. This number is proportional to time interval dt and to total number N of radioactive nuclei:

$$dN = -\lambda N\,dt, \tag{32.8}$$

where λ is ***decay constant***, which is proportional to the probability of decay of a radioactive nucleus and is different for different radioactive materials. The negative sign is put because dN <0 since the number of radioactive nuclei decreases over time.

Separating the variables and integrating (32.8), taking into account that the lower integration limits correspond to the initial conditions ($t = 0$, $N = N_0$, where N_0 is the initial number of radioactive nuclei), and high limits are equal to the actual values of t and N: $\int\limits_{N_0}^{N} \frac{dN}{N} = -\lambda \int\limits_{0}^{t} dt$, $\ln\frac{N}{N_0} = -\lambda t$. Finding the antilogarithm of this expression, we obtain:

$$N = N_0 e^{-\lambda t}. \tag{32.9}$$

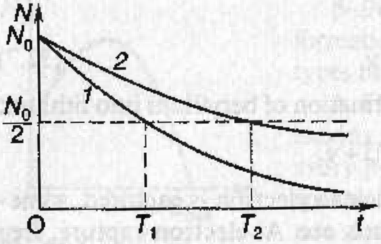


Fig. 32.2

This is the basic law of radioactive decay: the number of radioactive nuclei that have not yet decayed decreases over time at an exponential rate.

Fig. 32.2 shows curves *1* and *2* corresponding to different materials ($\lambda_1 > \lambda_2$); initial number N_0 of radioactive nuclei is the same.

In practice, instead of a decay constant, another characteristic of a radioactive isotope is often used: ***half-life*** *T*. It is the time during which a half of radioactive nuclei decays. Naturally this definition is valid for a sufficiently large number of nuclei. Fig. 32.2 shows how using curves *1* and *2*, it is possible to find half-lives of nuclei; the straight line corresponding to $N_0/2$ is drawn to intersection points with the curves. The abscissas of the intersections points give T_1 and T_2.

To determine the relation between T and λ, we substitute $N = N_0/2$ and $t = T$. $N_0/2 = N_0 e^{-\lambda T}$ into equation (32.9). Canceling N_0 and finding the natural logarithm of this equation, we obtain:

$$T = \ln 2/\lambda \approx 0.69/\lambda. \qquad (32.10)$$

Working with radioactive sources it is important to know the number of particles or γ-photons leaving the specimen per second. This number is proportional to the decay rate, therefore it is referred to as ***activity***, an essential characteristic of a radioactive sample:

$$\mathrm{A} = \frac{\mathrm{d}N}{\mathrm{d}t}. \qquad (32.11)$$

Using (32.8)–(32.10), it is possible to find the following dependences for an activity:

$$\mathrm{A} = \frac{\mathrm{d}N}{\mathrm{d}t} = \lambda N = \lambda N_0 e^{-\lambda t}, \qquad (32.12)$$

$$\mathrm{A} = \frac{N}{T}\ln 2. \qquad (32.13)$$

Thus, activity of a sample is the higher the more radioactive nuclei there are, and the less their half-life of decay is. The sample activity decreases exponentially over time.

The unit of activity is a *becquerel* (Bq), it corresponds to such an activity of a nuclide in a radioactive source, in which one act of decay occurs per 1 second. The most common unit of activity is curie (Ci); 1 Ci = $3.7 \cdot 10^{10}$ Bq =

$= 3.7 \cdot 10^{10}\ s^{-1}$. Besides, there is another non-SI unit of activity – *rutherford* (Rd); $1\ Rd = 10^6\ Bq = 10^6\ s^{-1}$. To characterize the activity of the unit mass of a radioactive source, there is a quantity called ***unit mass activity*** equal to the ratio of isotope activity to its mass. Unit mass activity is expressed in ***becquerels per kilogram*** (Bq/kg).

§ 32.3. INTERACTION OF IONIZING RADIATION WITH MATERIALS

Charged particles and γ-photons propagating in a material interact with electrons and nuclei; the result is certain changes in the state of both the material and particles.

Upon passing through the material, the principal mechanism of energy loss by a charged particle (alpha and beta ones) is ionizing deceleration. In this case, its kinetic energy is expended for excitation and ionizing of atoms of the medium.

Linear specific ionization, linear stopping power, and average linear particle path of a material quantify the interaction of a particle with a material.

By ***linear specific ionization*** i we mean the ratio between the number dn of ion pairs formed by a charged ionizing particle over elementary path dl, and this path: $i = dn/dl$. The dimension is ion pairs/m.

Linear stopping power S of a material is the ratio between energy loss dE of a charged ionizing particle over elementary path dl in the material, and the length of this path: $S = dE/dl$. The dimension is J/m.

Mean linear range R of a charged particle is the average path passed by a particle into a given material until the particle loses its ionizing power.

The curve of a linear specific ionization dependence on path x passed by an alpha particle in a medium (air) is shown in fig. 32.3. As the particle is moving in a medium, its energy and velocity decreases. Linear specific ionization increases and decreases dramatically only at the end of the moving. The increase of i is due to the fact that at a lower velocity, the alpha particle stays longer close to an atom and thus, the probability of the atom ionization increases. As we can see in the figure, in the air at normal pressure, the linear specific ionization of alpha particles of natural radioactive isotopes is $i = (2 \div 8) \cdot 10^6$ ion pairs/m.

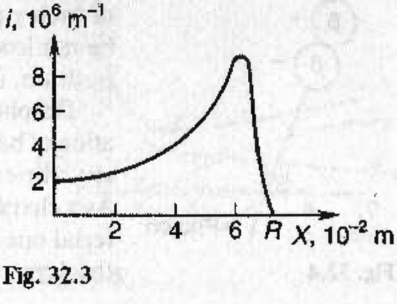


Fig. 32.3

Since ionization of one molecule takes the energy of about 34 eV, the values of the linear stopping power S of the material (air) is in the range of 70–270 MeV/m.

The mean linear range of an alpha particle depends on the energy of the particle. In the air, it is equal to several centimeters, in liquids and a living organisms, it is 10–100 μm. After the velocity of an alpha particle decelerat to the velocity of thermal motion of molecules, an alpha particle, having captured two electrons in the material, transforms into a helium atom.

Ionization and excitation are primary processes. Secondary processes can be as follows: acceleration of thermal motion of molecules, characteristi X-radiation, radioluminescence, chemical processes.

Interaction of alpha particles with nuclei is a much more rare process than ionization. In this case, nuclear reactions are possible, as well as scattering of alpha particles.

β-Radiation, as well as alpha radiation, causes ionization of a substance. In the air, linear specific ionization of β-particles can be calculated by the formula:

$$i = k(c/\upsilon^2),$$

where $k \approx 4600$ ion pairs/m.

In addition to ionization and excitation, beta particles can also induce other phenomena. For example, when electrons are decelerated, they generate bremsstrahlung. β-Particles are scattered by the electrons of a material, their trajectories curve in it very much. If an electron moves in a medium at a speed exceeding the speed of light in this medium, ***characteristic Cherenkov radiation* occurs (Cherenkov radiation).**

When a β^+-particle gets into a material, with a high probability, the interaction with an electron will produce two gamma photons rather than an electron-positron pair. This process is called annihilation, its diagram is shown in fig. 32.4. The energy of each γ-photon occurring upon annihilation must be not lower than the rest energy of the electron or positron, i.e. not lower than 0.51 MeV.

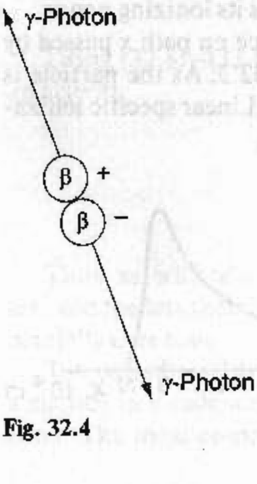


Fig. 32.4

Despite the variety of processes leading to attenuation of beta radiation, we can consider that its intensity varies at an exponential law rate like (31.8). As a characteristic of β-radiation absorption by a material one uses the half-value layer. Passing through this layer, the radiation intensity is reduced twice.

β-Particles are considered to penetrate body tissues 10–15 mm deep. Means of protection against β-radiation are thin aluminum, acrylic plastic, and other shields. For example, a layer of aluminum 0.4 mm thick or water 1.1 mm thick halves β-radiation from phosphorus $^{32}_{15}$P.

When γ-radiation gets into a material, along with processes characteristic for X-radiation (coherent scattering, Compton effect, photoelectric effect, see § 31.3), there are also phenomena that are not typical of interaction of X-rays with a material. These processes are induced by the electron–positron pair, when the γ-photon energy is not less than the total rest energy of electron and positron (1.02 MeV), and photonuclear reactions, occurring upon interaction of high energy gamma photons with atomic nuclei. For a photonuclear reaction to occur, the γ-photon energy should be not less than binding energy per nucleon. Because of various processes occurring under the impact of gamma radiation, charged particles are formed; therefore, γ-radiation is an ionizing radiation too.

In a substance, attenuation of γ-radiation is usually described by an exponential function (31.8). Linear (or mass) absorption factor can be represented as a sum of the corresponding absorption factors taking into account three principal processes of interaction, i.e. photoelectric effect, Compton effect and formation of electron-positron pairs:

$$\mu = \mu_{ph} + \mu_{ce} + \mu_p. \qquad (32.14)$$

These principal processes of interaction take place with different probability which depends on the γ-photon energy (fig. 32.5; the curve is obtained for lead). As we can see in the figure, at low energies, the photoelectric effect is the most important; at medium energies, it is the Compton effect, and at energies higher than 10 MeV, it is electron–positron pair formation.

The exponential function of gamma photon beam attenuation is valid approximately, especially at high energies. This is due to secondary processes that take place upon interaction of γ-radiation with the material. For example, electrons and positrons have energy, which is sufficient to form new γ-photons due to deceleration and annihilation.

A neutron beam is also ionizing radiation since interaction of neutrons with the nuclei of atoms produces charged particles and γ-radiation. Let us illustrate this with several examples.

- Nuclear fission when nuclei capture neutrons forms radioactive fragments, γ-radiation and charged particles.

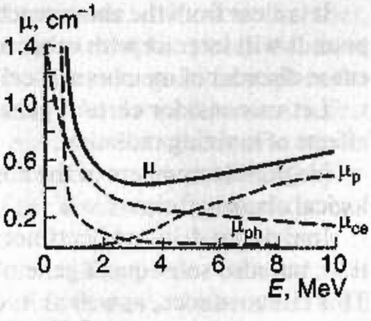


Fig. 32.5

- Formation of α-particles, for example, ${}^{27}_{13}\mathrm{Al} + {}^{1}_{0}n = {}^{24}_{11}\mathrm{Na} + {}^{4}_{2}\alpha$.
- Formation of protons, for example, ${}^{14}_{7}\mathrm{N} + {}^{1}_{0}n = {}^{14}_{0}\mathrm{C} + {}^{1}_{1}p$.

§ 32.4. BIOPHYSICAL PRINCIPLES OF IONIZING RADIATION EFFECT ON A BODY

Considering the primary physical and chemical processes in the body under the impact of ionizing radiation, it is necessary to take into account two fundamentally different possibilities of interaction: with water molecules and molecules of organic compounds.

Under the impact of ionizing radiation, there are chemical transformations of the material called *radiolysis*. Let us enumerate the possible mechanisms of water *radiolysis*:

$$\mathrm{H_2O} \to \mathrm{H_2O^*}, \quad \mathrm{H_2O^-} \to \mathrm{OH^-} + \mathrm{H}.,$$
$$\mathrm{H_2O} \to \mathrm{H_2O^+} + e^-, \quad \mathrm{OH^-} \to \dot{\mathrm{O}}\mathrm{H} + e^-,$$
$$\mathrm{H_2O^+} + \mathrm{H_2O} \to \mathrm{H_3O^+} + \mathrm{O}\dot{\mathrm{H}}, \quad \mathrm{H^+} + e^- \to \dot{\mathrm{H}},$$
$$\mathrm{H_2O} + e^- \to \mathrm{H_2O^-}.$$

Interaction with oxygen can lead to formation of hydrogen hydroperoxide and peroxide:

$$\dot{\mathrm{H}} + \mathrm{O_2} + \mathrm{HO_2}, \quad \mathrm{HO_2} + \dot{\mathrm{H}} \to \mathrm{H_2O_2}.$$

Interaction of ionizing radiation with molecules of organic compounds can form excited molecules, ions, radicals, and peroxides:

$$\mathrm{RH} \to \mathrm{RH^*} \to \mathrm{R} + \mathrm{H}, \quad \mathrm{RH^+} \to \mathrm{R} + \mathrm{H^+},$$
$$\mathrm{RH} \to \mathrm{RH^+} + e^-, \quad \mathrm{R} + \mathrm{O_2} \to \dot{\mathrm{R}}\mathrm{O_2}.$$

It is clear from the above reactions that these chemically highly active compounds will interact with other molecules of the biological system, which will cause disorder of membranes, cells, and functions of the entire organism.

Let us consider certain general regularities characteristic for biological effects of ionizing radiation.

Negligible numbers of the absorbed radiation energy can cause grave biological abnormalities.

Ionizing radiation affects not only the biological object exposed to irradiation, but also subsequent generations through the hereditary system of cells. This circumstance, as well as its conditional forecasting, raises the question of protection of organisms against radiation.

The *latent* period is specific for the biological effect of ionizing radiation. Different parts of cells show different sensitivity to the same dose (see Chapter 33) of ionizing radiation. The most sensitive is the nucleus of a cell.

The ability to divide is the most vulnerable cellular function, which is why irradiation affects the growing tissues primarily. This makes ionizing radiation especially dangerous for children, including the the time in utero. Radiation is destructive for the tissues of an adult body, in which there is regular or periodic cell division: gastric and intestinal mucosa, hematopoietic tissue, gametes, etc. The action of ionizing radiation on fast-growing tissues is also used in therapy for malignant tissues.

At high doses, a lethal acute radiation syndrome can occur, at lower doses, various diseases develop like radiation syndrome, etc.

§ 32.5. IONIZING RADIATION DETECTORS

Ionizing radiation detectors are instruments recording α-, β-diation and γ-radiation, neutrons, protons, etc. These devices are also used to measure the energy of particles, to study the processes of their interaction, decay, etc.

Operation of detectors is based on processes caused by particles detected in a material.

With certain conventionality, detectors can be subdivided into three groups: ***track*** (trace) detectors, ***counters***, and **integrated *instruments*.**

Track detectors make it possible to observe the trajectory of a particle, counters record the presence of a particle in a given space, and integrated instruments give information about the flow of ionizing radiation. Once again, this classification is quite conventional. Thus, for example, track detectors can be used to count the passing particles; from single-piece recordings of particles by a counter, it is possible to pass to total estimation of ionizing radiation flow, etc.

Track detectors include a Wilson cloud chamber, diffusion chamber, bubble chamber, spark detector and thick-layer photographic plates. All these devices share the common principle that the observed a particle ionizes the molecules or atoms of the material in its path. The formed ions manifest themselves through their secondary effects: *condensation of a supersaturated* **vapor** (Wilson cloud chamber and diffusion chamber); ***vaporization of superheated liquid*** (bubble chamber); ***formation of gas discharges*** (spark detector); ***photochemical action*** (thick-layer photo plates). Since the reader knows many of these methods from the secondary school curriculum, as an illustration, let us consider only operation of a spark chamber. It consists of electrodes with a gas between them. High voltage is applied at the electrodes when a particle

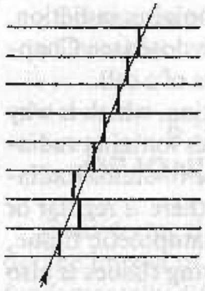


Fig. 32.6

passes through the chamber space; the signal to turn on power comes from other detectors. The electrons, which appear along the trajectory of the particle due to ionization of gas atoms, are accelerated by the electric field and produce collisional ionization themselves. As a result, a visible spark discharge is formed in small parts of the trajectory.

Fig. 32.6 shows a diagram of a narrow-gap spark chamber. The distance between the electrodes placed in the chamber is about 1 cm. Spark discharges, which occur, are perpendicular to the electrodes, their totality shows the trajectory of the particle.

In the streamer[1] chamber, the distance between the electrodes is 5–20 cm. The high voltage is removed approximately 10^{-5} s after passage of the particle. During this time, sparks are generated only in the immediate region of primary ionization induced by the particle recorded. Traces of particles of a streamer chamber are shown in fig. 32.7.

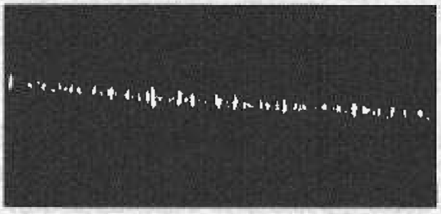


Fig. 32.7

Integrated detectors include photographic films (after the film development, the degree of blackening is recorded), continuous operation ionization, chambers, etc.

Let us consider the structure and operation of the continuous operation ionization chamber. It has capacitor C with a gas inside (fig. 32.8). When radiation enters the chamber, ionization occurs, and an electric current flows through the circuit; the current is usually amplified and measured. The current intensity is proportional to the number of ions generated in the chamber per second and, therefore, to the energy flow of the passing ionizing particles.

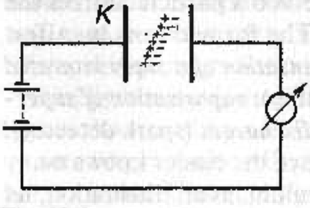


Fig. 32.8

In certain instruments, discharge of the capacitor under the impact of ionizing radiation is recorded by an electrometer.

---

[1] *Streamers* are luminous branched channels formed by electrical discharge in gases.

Counters include a large group of gas-discharge instruments (pulse ionization chambers, proportional counters, and Geiger–Müller counters) as well as fluorescent, semiconductor counters, etc.

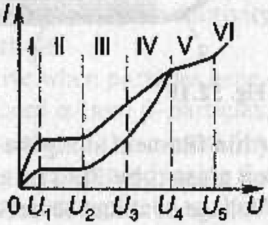


Fig. 32.9

Let us analyze the dependence of current pulse intensity I, which occurs when a particle enters the gas gap (the number of ions participating in one pulse), on voltage U at electrodes (fig. 32.9; the curves correspond to α- and β-particles).

Both curves can be conventionally represented by six areas, which are characterized by different processes.

In recombination area I, a part of ions recombines. As the voltage is increasing, the number of recombining ions decreases, and the number of ions that reach the electrodes, increases. As the ionizing power of α-particles is higher than that of β-particles, their curves are different.

Area II corresponds to saturation. All primary ions reach electrodes, but there is no secondary ionization yet. The ionizing chamber operates in this area.

In region III, secondary ionization begins to become manifest, but the current pulse remains proportional to initial ionization. After amplification, the number of ion pairs N is proportional to the number of primary ion pairs N_0 formed by the ionizing particle:

$$N = kN_0, \tag{32.15}$$

where k is the coefficient of gas amplification ($k = 10^3 \div 10^6$). It depends on the counter structure and the nature of gas used in it. It is in this particular area that proportional counters operate.

Since N_0 and, therefore, N depend not only on the type of particle, but also on its energy, proportional counters can measure the energy of particles.

Area IV is called area of limited proportionality. There is still a dependence on primary ionization, but if the value is equal to U_4, it has already vanished. Value U_4 is called Geiger threshold; it depends on the counter design, as well as the pressure and the type of gas used in the counter. In this area, the current pulse becomes sufficiently high at low primary ionization too.

Geiger–Müller counters operate in area V. Here the coefficient of gas amplification is high, but it is impossible to distinguish the energies of particles.

In area VI, a continuous gas discharge occurs, which leads to quick damage of the counter. Areas V and VI correspond to a self-maintaining gas discharge, which is maintained after the ionizing action of particles has stopped.

As an example of gas instruments, let us consider a ***Geiger–Müller counter***, it consists of coaxially arranged cylindrical electrodes [fig. 32.10: *1* is an anode

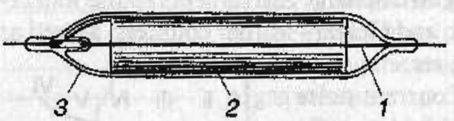


Fig. 32.10

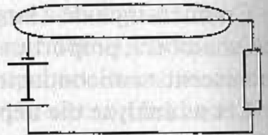


Fig. 32.11

(thin filament along the axis), *2* is a cathode in the form of metal spray coating on glass tube *3*]. The gas pressure inside the counter is 100–200 mm Hg. Voltage of abiout several hundred V is applied to the electrodes.

When an ionizing particle enters the counter, free electrons are formed in the gas and move to the anode. Since the filament is thin (its diameter is about 0.05 mm), near it, the electric field is highly heterogeneous, the field intensity is high. Near the filament, electrons are accelerated so that they begin to ionize the gas. As a result, there is a discharge and a current flows through the circuit (fig. 32.11).

An independent discharge in a Geiger–Müller counter should be quenched; otherwise, the counter will not react to the next particle. To quench the discharge, the radio engineering method and one based on addition of polyatomic gases in the tube (self-quenched counters) are used.

The simplest variant of the first method is connecting a high-ohm resistor in series with the counter. When a current flows, there is a sizeable voltage drop across this resistor, the voltage at the counter decreases, and the discharge stops. Self-quenched counters are more common: due to a special gas inside, the discharge cuts off by itself even at low circuit resistances.

Electrical pulses occurring in the external circuit at the resistor are amplified and recorded by a special device. Fig. 32.12 shows the appearance of installation B-4, which operates jointly with a Geiger–Müller counter.

The operational principle of the scintillation (luminescence) counter is due to short flashes, which are called scintillations, arising in some materials under the impact of ionizing radiation. At the first stage of nuclear physics

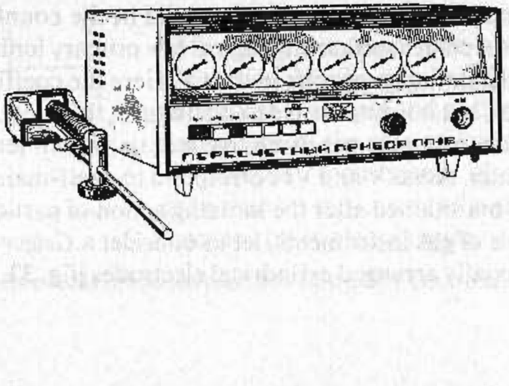


Fig. 32.12

development, scintillations were recorded visually. In the luminescence counter, they are recorded automatically with a photoelectronic multiplier.

Semiconductor counters react to changes in the electrical conductivity of the *p*–*n*-junction under the impact of a charged particle.

As you can see, all the foregoing detectors are active when particles generate ionization of a certain volume. Therefore, to record α- and β-particles, the walls of chambers and counters should be transparent to these particles. Sometimes, to record α-radiation, a corresponding source is placed inside the chamber since it is difficult to make the chamber walls transparent to these particles.

X-ray and γ-radiation are recorded due to ionization caused by charged particles generated by the photoelectric effect, Compton effect, etc.

The counters should meet certain general requirements, such as efficiency, resolution time, etc. ***Efficiency*** is the ratio between the amount of registered particles to the total number of particles passing through the counter. ***Resolution*** (or *dead*) ***time*** of a counter is a minimum time interval separating successive particles so that they are registered separately.

§ 32.6. USE OF RADIONUCLIDES AND NEUTRONS IN MEDICINE

Medical applications of radionuclides can be classified as two groups. One group includes methods that use radioactive indicators (labeled atoms) for diagnosis and research. The other group of methods is based on the use of ionizing radiation of radionuclides for a biological therapeutic effect. Bactericidal action of radiation can be attributed to the same group.

The method of labeled atoms is delivering radionuclides into the body and determining of their location and activity in organs and tissues. For example, to diagnose a thyroid gland disease, radioactive iodine $^{125}_{53}\mathrm{I}$ or $^{131}_{53}\mathrm{I}$ is administered, a part of this material is concentrated in this gland. A counter, placed close to it, registers the amount of accumulated iodine. The rate of increase in the concentration of radioactive iodine makes it possible to make a diagnostic conclusion of the actual state of the thyroid gland.

Thyroid gland cancer can metastasize into different organs. Data about accumulation of radioactive iodine in them can give information about metastases.

To detect distribution of radionuclides in different organs of the body, a *scintiscanner* is used; it automatically records the distribution and intensity of a radiopharmaceutical. A scintiscanner is a scanning counter, which gradually passes over large areas of the patient's body. A single scintillation because of radiation is recorded, for example, by a bar stroke on paper. Fig. 32.13, a shows sketchily the passage of a counter, and fig. 32.13, **b** shows its recording card.

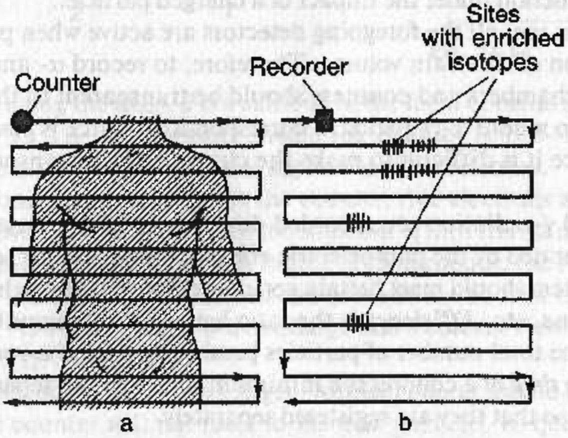


Fig. 32.13

Using radioactive indicators, it is possible to trace metabolism in the body. It is difficult to measure the volume of body fluids directly; the method of labeled atoms makes it possible to solve this problem. For example, administer a certain amount of radioactive indicator into the blood and waiting for its homogeneous distribution through the blood circulatory system, it is possible to find the total volume of the indicator by the activity of the unit of blood volume.

A scintiscanner yields a relatively rough distribution of ionizing radiation in organs. More detailed information can be obtained by the ***autoradiography*** method.

In this method, the test object, such as a biological tissue, is covered with a layer of sensitive emulsion. Radionuclides in the object leave a trace in the corresponding place of the emulsion, as if taking their own photo (hence the name of the method). The obtained image is called a ***radioautograph*** or ***autoradiogram***. Fig. 32.14 illustrates application of this method. Here is an image of a rat thyroid follicle (a) and an autoradiogram (b) of the same follicle after administration of an amino acid (leucine) labeled with radioactive ^{14}C. Developed grains of silver (dark spots in photographic emulsion in fig. 32.15) indicate the distribution of ^{14}C in the follicle.

Radioactive atoms are delivered into a living organism in such a small anount that neither they nor their decay products cause any detriment to the body.

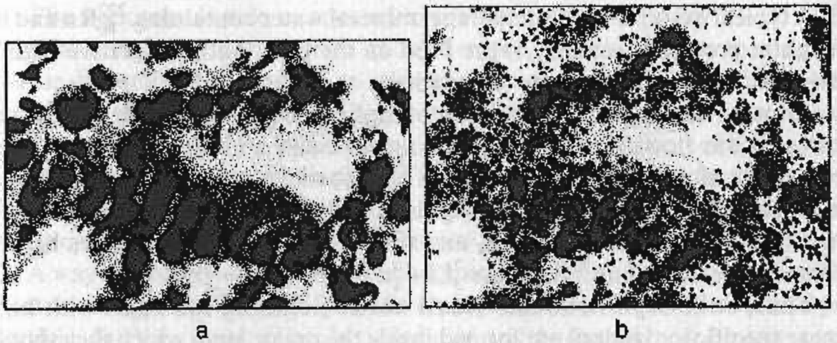


Fig. 32.14

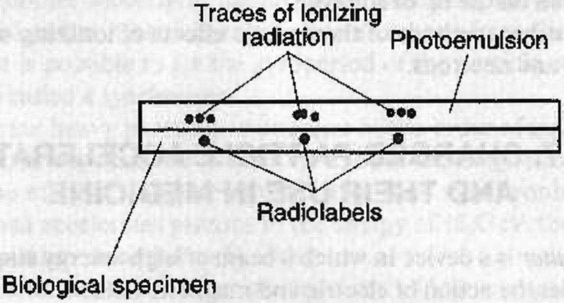


Fig. 32.15

Therapeutic use of radionuclides is mostly connected with γ-radiation (gamma-therapy).

A gamma ray source (fig. 32.16) consists of the source proper (usually, it is ^{60}Co), and a protective container, inside which the source is placed; the patient is placed on a table. Administration of γ-radiation of high energy makes it possible to destroy deeply located tumors while superficial organs and tissues are exposed to less destructive action.

α-Particles have a therapeutic use too. Since they have sizeable linear specific ionization, even a small layer of the air absorbs it. Therefore, in therapy, use of α-particles (α-therapy) is only possible upon a contact with the body or internal administration.

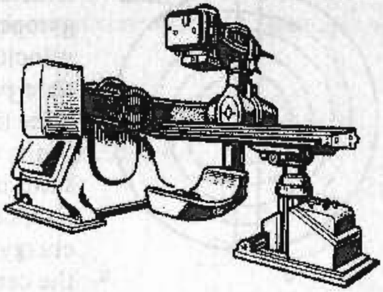


Fig. 32.16

A typical example is radon therapy: mineral water containing $^{222}_{86}Rn$ and its daughter products (see § 32.1) are used on the skin (bath), digestive organs (drinking), respiratory system (inhalation).

Another therapeutic application of alpha particles is mostly connected with neutron flows. Elements, whose nuclei cause a nuclear reaction under the impact of neutrons with formation of α-particles, are delivered into a tumor mass. After that, by irradiating the affected organ with a neutron flow, one induces a nuclear reaction, and thus formation of α-particles begins (for example, reactions $^{10}_{5}B + ^{1}_{0}n \to ^{7}_{3}Li + ^{4}_{2}\alpha$ or $^{6}_{3}Li + ^{1}_{0}n \to ^{3}_{1}H + ^{4}_{2}\alpha$).

Thus, both α-particles and recoil nuclei (ionizing radiation with high linear specific ionization) are formed inside the organ, onto which they should exert their damaging action. It is possible to administer a radioactive material into a bad organ on the tip of a needle.

There are other methods of therapeutic effects of ionizing radiation from radionuclides and neutrons.

§ 32.7. CHARGED PARTICLE ACCELERATORS AND THEIR USE IN MEDICINE

An accelerator is a device in which a beam of high-energy charged particles is formed under the action of electric and magnetic field.

There are *linear* and *cyclic accelerators*. In linear accelerators, particles move along a rectilinear trajectory; in cyclic ones, they move along a circle or spiral.

The most known cyclic accelerator is a *cyclotron* (fig. 32.17), in which a charged particle moves along a circumference under the action of a magnetic field with induction B. Its direction is perpendicular to the plane of the figure. The alternating electric field between dees *1* accelerates the particle. According to formula (16.28), gyroperiod *T* of the particle does not depend on its velocity or the radius of its trajectory, so the particle passes any semi-circumference within the same time. This time is a half an oscillation period of the electric field. Thus, the magnetic field ensures movement of the particle along the circumference, and the electric field changes the kinetic energy of the particle. Source *2* of particles is near the center of the cyclotron, beam *3* of accelerated particles leaves the cyclotron after acceleration.

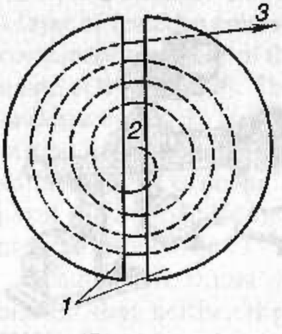


Fig. 32.17

The cyclotron is able to accelerate protons to the energy of 20–25 MeV. Limitation of the energy value of accelerated particles is due to the relativistic dependence of mass on its velocity. Since the mass increases in proportion to its velocity [see (16.28)], the gyroperiod of the particle will increase too. As a result, there will be no synchronism between the motion of the particle and the change in the electric field. The electric field will not accelerate, but decelerate the particles. That is why, it is impossible to accelerate electrons in the cyclotron, since they attain relativistic velocities quickly.

A way out of this difficulty can be found by varying the frequency of the electric field in accordance with the change in the gyroperiod of a charged particle. Such an accelerator is called a *phasotron* (*synchrocyclotron*); it can accelerate protons to the energy of 730 MeV.

There is another solution to the problem: increasing the induction of the magnetic field as the mass increases. As we can see from formula (16.28), in this case, it is possible to fix the gyroperiod of the particle. An accelerator of this type is called a *synchrotron*.

To accelerate heavy particles to energies of the order of gigaelectronvolts and more, a *synchrophasotron* is used; in it, both the magnetic field and frequency of the electric field are alternating. The synchrophasotron operating in Dubna accelerates protons to the energy of 10 GeV, the synchrophasotron in Serpukhov does this to 76 GeV.

A quite common accelerator of electrons of low energy is a ***betatron.*** In contrast to other cyclic accelerators, the electric field is not supplied from external sources, but is generated due to magnetic field changes.

Fig. 32.18, a, illustrates how, upon a change in the magnetic field of the electromagnet *1*, according to Maxwell's theory, an eddy electric field is generated. In gap *2* of the magnet, there is a vacuum chamber, in which electrons are accelerated. The force lines of the electric field in the form of concentric circumferences are located in a plane perpendicular to the plane of fig. 32.18, a. Fig. 32.18, b, shows an individual line of electric field intensity, which approximately coincides with the electron trajectory. In this figure, the lines of vector $\vec{B}$ are mainly perpendicular to the plane of the drawing, the magnetic induction increases.

The electron is held in its orbit by a magnetic field (the Lorentz force) and is accelerated by the electric field.

Betatrons can accelerate electrons to tens of megaelectronvolts. Nowadays, betatrons find many applications including

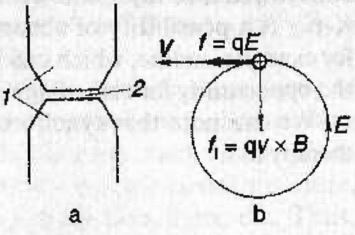


Fig. 32.18

medicine. Let us dwell on medical application of accelerators. Accelerato of charged particles are used as a means of radiothrapy in two main directio

First, bremsstrahlung is used; it occurs when electrons accelerated are decelerated by a betatron. A photon of bremsstrahlung with energy of several tens of MeV is more efficient than γ-radiation of most isotopes.

Second, the direct action of accelerated particles is used: electrons, protons. Electrons are accelerated by a betatron, and a proton beam is obtained by other accelerators. As we can see in fig. 32.3, charged particles, including protons, cause the highest ionization before they stop. Therefore, when a proton beam enters a biological object from without, the greatest impact will not be on the surface layers, but on tumor tissues located deep in the body. This is the main advantage of charged particles use for radiotherapy of deeply located tumors. The surface layers are affected minimally in this case. Low scattering of protons makes it possible to form narrow beams and, therefore, affect precisely the tumor. In recent years, beside therapeutic use of accelerators, their diagnostic use has become prominent. We can consider the following two spheres in this.

One area is *medical ion radiography*. The essence of the method is the following. The path of heavy charged particles (α-particles, protons) depends on the absorbing material density. Therefore, recording the flow of particles before and after passage through the object makes it possible to obtain information about the average density of the material.

Thus, like in X-ray radiography, it is possible to distinguish structures of higher and lower density. The advantage of this method over X-ray radiography is a lower contrast, which makes it possible to distinguish the structure of soft tissues better.

Another area of application is synchrotron radiation.

Synchrotron radiation is intensive ultraviolet and soft X-rays emitted by electrons moving along a circular orbit with velocities close to the speed of light. This radiation was first observed as light radiation in synchrotrons, that is why it is called synchrotron radiation. It is used for diagnostic purposes like conventional X-rays. One of the advantages of synchrotron radiation over X-ray is a possibility of absorption of this radiation by such elements as, for example, iodine, which can have excessive concentration in tissues. Hence the opportunity for early diagnostics of malignant tumors.

We can note that synchrotron radiation begins to be used in radiation therapy too.

Chapter 33

Introduction to dosimetry. Elementary particles

The need to quantify the effect of ionizing radiation on various materials of living and inanimate nature has led to origin of dosimetry. Dosimetry is a branch of nuclear physics and measurement technology studying values characterizing the action of ionizing radiation on materials as well as methods and instruments for their measurement. Initially, development of dosimetry was due to the need to take into account the effect of X-rays on people.

§ 33.1. RADIATION AND EXPOSURE DOSE. DOSE POWER

It was noted before that ionizing radiation has an action on a material only when it interacts with particles of the chemical composition of that material.

Irrespective of the nature of ionizing radiation, its interaction can be quantified by the ratio between the energy imparted to the element of the irradiated material, and the mass of that element. This characteristic is called ***radiation dose (absorbed radiation dose) D.***

The various after-effects of ionizing radiation are first determined by the absorbed dose of radiation. It has a complex dependence on the type of ionizing radiation, energy of its particles, and chemical composition of the irradiated material, and is proportion to the time of irradiation. The relation between the dose and time is called ***dose rate.***

The unit of absorbed radiation dose is *gray* (Gy), which corresponds to a radiation dose, at which the irradiated material mass of 1 kg receives the ionizing radiation energy of 1 J; *radiation dose rate* is expressed in *grays per second* (Gy/s). A non-SI unit of radiation dose is *rad*[1] (1 rad = 10^{-2} Gy = = 100 erg/g), a dose rate is *rad per second* (rad/s).

It would seem that to find the absorbed radiation dose it is necessary to measure the energy of ionizing radiation incident on the body, energy passing through the body, and divide their difference by the body mass. Nevertheless, in practice, this is difficult to do, since a body is a heterogeneous structure, the imparted energy is scattered by the body in all directions, etc. Thus, a quite specific and clear concept of a radiation dose is not of great use in experiments. However, it is possible to estimate the absorbed dose by the ionizing effect of radiation in the air surrounding the body.

---

[1] The term rad is the abbreviation of the following words: Radiation Absorbed Dose.

In this connection, another concept of dose for X-ray and γ-radiation is introduced: ***exposure dose of radiation X***, which is a measure of air ionization by X-rays and γ-rays.

The unit of exposure dose is a *coulomb per kilogram* (C/kg). In practice, a unit called *Roentgen* (R) is used: this is exposure dose of X-rays or gamma radiation at which due to total ionization, in one cm^3 of dry air (at 0 °C and 760 mm of Hg), ions are formed carrying a charge of 1 unit CGS charge of each sign. It is not difficult to calculate that the exposure dose of 1 R corresponds to formation of $2.08 \cdot 10^9$ ion pairs in 0.001293 g of dry air; $1\ \mathrm{R} = 2.58 \cdot 10^{-4}$ C/kg.

The unit of exposure dose rate is 1 A/kg, and a non-SI unit is 1 R/s.

Since the radiation dose is proportional to the incident ionizing radiation, there should be a direct proportional relationship between the radiation dose and exposure dose:

$$D = fX, \tag{33.1}$$

where f is a certain transition coefficient depending on a number of factors and, first of all, on the irradiated material and photon energy.

The easiest way to determine the value of coefficient f is when the irradiated material is the air. If $X = 1$ R, in 0.001293 g of air, $2.08 \cdot 10^9$ ion pairs are formed; therefore, in 1 g of air, there are $2.08 \cdot 10^9/0.001293$ ion pairs. On average, for formation of one pair of ions, it takes the energy of 34 eV. This means that 1 g of air absorbs a radiation energy equal to:

$$\frac{2.08 \cdot 10^9}{0.001293} 34 \cdot 1.6 \cdot 10^{-12}\ \mathrm{erg/g} = 88\ \mathrm{erg/g}.$$

Thus, in the air, from the energy viewpoint, the absorbed dose of 88 erg/g is equivalent to 1 R. Then, according to formula (33.1), we have:

$$D = 0.88X,\ \ f = 0.88,$$

if D is measured in rads and X is measured in R.

For the air, coefficient f does not strongly depend on the energy of photons.

For water and soft tissues of the human body, $f = 1$; therefore, in rads, the radiation dose is numerically equal to the corresponding exposure dose in R. This to be exact is the reason to use non-SI units, which are rad and R.

For osseous tissue, coefficient f decreases when the energy of photon increases from about 4.5 to 1.

Let us establish a connection between the activity of a radioactive material —the source of gamma photons — and the rate of exposure dose. Gamma photons leave source U (fig. 33.1) in all directions. The number of these photons passing through 1 m^2 of a certain sphere surface per 1 s, is directly proportional to activity A and inversely proportional to the area of the spherical surface ($4\pi r^2$). In volume V, the exposure dose rate (X/t) depends on this number of photons, since it is they who induce ionization. Whence we obtain:

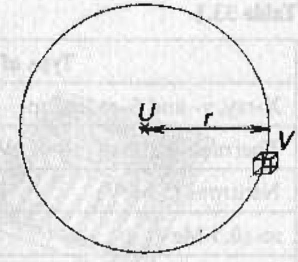


Fig. 33.1

$$\frac{X}{t} = k_\gamma \frac{A}{r^2}, \qquad (33.2)$$

where k_γ is the specific gamma-ray dose constant that is typical of a given radionuclide.

§ 33.2. QUANTITATIVE ESTIMATION OF BIOLOGICAL EFFECTS OF IONIZING RADIATION. EQUIVALENT DOSE

For a given type of radiation, the biological effect is usually the greater the higher the dose of radiation is. However, different types of radiation, even with the same absorbed dose, produce different effects.

In dosimetry, it is accepted to compare biological effects of different types of radiation with the corresponding effects caused by X-rays and γ-radiation.

The coefficient K, showing how many times the effectiveness of the biological effect of this radiation type is greater than the effectiveness of X-ray or γ-radiation upon the same radiation exposure in the tissues, is a ***quality factor***. In radio-biology it is also called relative ***biological effectiveness*** (RBE).

The quality factor is determined on the basis of experimental data. It depends not only on the type of particle but also on its energy. Let us give approximate values of K (Table. 33.1) for certain types of radiation (the energy of particles is indicated in brackets).

An absorbed dose together with quality factor gives an idea of biological effect of ionizing radiation, so product DK is used a universal measure of this effect and is called equivalent radiation exposure H:

$$H = DK. \qquad (33.3)$$

Table 33.1

Type of radiation	K
X-ray, γ- and β-radiation	1
Thermal neutrons (~0.01 eV)	3
Neutrons (5 MeV)	7
>> (0.5 MeV)	10
α-Radiation	20

Since K is a dimensionless coefficient, an equivalent radiation dose has the same dimension as an absorbed radiation dose, but it is called sievert (Sv). The non-SI unit of equivalent dose is rem[1]; 1 rem = 10^{-2} Sv.

In rems, the equivalent dose is equal to the radiation dose in rads multiplied by the quality factor.

Natural radioactive sources (cosmic rays, radioactivity of the earth depths, water, nuclei that make up the human body, etc.) produce background radiation of approximate equivalent dose of 125 mrem. The maximum permissible dose equivalent for occupational exposure is considered to be 5 rem annually. The minimum lethal dose for γ-radiation is about 600 rems. This data correspond to irradiation of the whole body.

§ 33.3. DOSIMETRY INSTRUMENTS

Dosimetry instruments, or ***dosimeters***, are instruments for measuring doses of ionizing radiation or quantities connected with these doses.

Structurally, dosimeters consist of a nuclear radiation detector and measuring instrument. They are usually graduated in units of dose or dose rate. In certain cases, an alarm signal of exceeding the set quantity of the dose rate is designed.

Depending on the detector used, there are ionizing, luminescent, semiconductor, photo dosimeters, etc.

Dosimeters can be designed to measure doses of a defined type of radiation or to record mixed radiation.

Devices for measuring the exposure dose of X-ray and γ-radiation or the dose rateare called ***X-ray intensity meters*** (R-meters).

As a detector, they usually use an ionization chamber. The charge flowing through the chamber circuit is proportional to the exposure dose, and the current intensity is proportional to its power. Fig. 33.2 shows a micro R-meter

---

[1] Rem is the abbreviation of "Roentgen equivalent man".

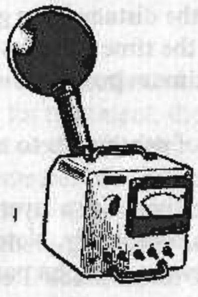


Fig. 33.2

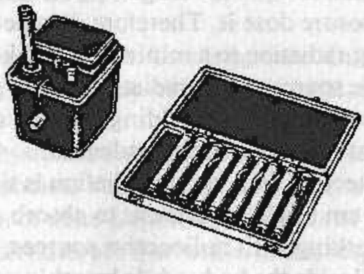


Fig. 33.3

MRM-2 with a spherical ionizing chamber located separately from the instrument.

The gas composition in the ionization chambers, as well as the material of their walls, are chosen to create such conditions for absorption of energy as exist in biological tissues.

Fig. 33.3 shows a set of individual dosimeters DK-0.2 with a common measuring instrument. Each individual dosimeter is a small cylindrical ionizing chamber that is pre-charged. Because of ionization, the chamber is discharged, which is read by a built-in electrometer. Its readings depend on the exposure dose of ionizing radiation.

There are dosimeters with gas discharge counters as detectors.

Instruments called ***radiometers*** are used to measure the activity or concentration of radioactive isotopes. The principle of their operation is described in § 32.5.

In conclusion, we will note that the general design of all dosimeters is like one shown in fig. 21.1. A nuclear radiation detector serves as a sensor (measuring transducer). As output devices, needle instruments, recorders, electromechanical counters, sound and light alarms, etc. can be used.

§ 33.4. PROTECTION AGAINST IONIZING RADIATION

Work with any sources of ionizing radiation requires protection of staff from their harmful action. This is a great and special problem, going beyond purely physical issues. Let us consider it briefly.

There are three types of protection: with time, distance, and shielding.

Let us illustrate the first two types of protection with a model of a point source of γ-radiation. Let us transform formula (33.2):

$$X = k_{\gamma} \frac{A}{r^2} t. \qquad (33.4)$$

We can see that the longer the time and the shorter the distance, the greater the exposure dose is. Therefore, it is necessary to cut the time of exposure to ionizing radiation to a minimum, and keep at the maximum possible distance from the source of this radiation.

The principle of shielding is based on the ability of substances to absorb different types of ionizing radiation.

Protection from alpha radiation is simple: a sheet of paper or a layer of air several cm thick is sufficient to absorb alpha particles completely. Nevertheless, working with radioactive sources, one should beware of alpha particles getting inside the body while breathing or eating.

For protection against β-radiation, it is sufficient to use plates from aluminum, acrylic glass or common glass several cm thick. Upon interaction of β-particles with a material, bremsstrahlung can occur; and β^+-particles can generate γ-radiation as they are annihilated by electrons. The most difficult protection is that against neutral radiation: X-ray and γ-radiation, neutrons. These types of radiation interact with the particles of a material with a lesser probability, therefore, penetrate deeper into the material. Attenuation of X-ray and γ-radiation approximately meets law (31.8). The attenuation factor depends on the number of the atomic number of the element of the absorber [see (31.12)] and wavelength, which for gamma photons, is shown in fig. 32.5. These dependencies, photon scattering and secondary processes are taken into account in calculating the parameters of protection. Some of them are shown for X-rays in fig. 31.10. Protection from neutrons is the most difficult. At first, fast neutrons are decelerated reducing their velocity in hydrogen-containing materials. After that, other materials, for example, cadmium, absorb slow neutrons.

CONCLUSION

Physical methods have been used in medicine for a long time. In ancient times, for treatment, they used cooling and heating various body parts, fixation of fractured limbs, etc.

A number of scientists, who were physicians and physiologists, in the course of their professional activities and hobbies, worked on physical issues thus strengthening by their works the fusion of these important branches of natural science. The examples of some great scientists are instructive in this respect.

Thomas Young (1773–1829) studied at several universities, where he at first studied medicine and then was attracted to physics. He explained the phenomenon of accommodation of the eye by changing the curvature of the lens; he was the first to explain the phenomenon of light interference and introduced the term "interference" itself, developed the theory of color vision, studied deformation of bodies.

Jean Louis Marie Poiseuille (1799–1869) was a French physicist and physiologist. He studied liquid flow in thin cylindrical tubes and internal friction and was the first to use a mercury manometer to take blood pressure.

Julius Robert von Mayer (1814–1878) was a German physician. As a naval surgeon, during navigation, he noticed that in the tropical and northern latitudes, the color of the venous blood of sailors differs. This gave brought him to a conclusion that there is a relationship between consumption of matter and formation of heat, and that heat and work can be mutually converted. Mayer was one of the first to discover the law of conservation and transformation of energy.

Hermann Ludwig Ferdinand von Helmholtz (1821–1894) was a German physician, physiologist, and physicist. He provided a mathematical demonstration to the law of conservation of energy noting its universal nature, developed the thermodynamic theory of chemical processes, achieved sizeable success in the field of physiological acoustics and physiology of vision, and measured the velocity of nervous excitation propagation for the first time.

Jeacques-Arsène d'Arsonval (1851–1940) was a French physicist and physiologist. He conducted research in the field of electricity and its application in medicine, was the founder of electro-physiotherapy.

In medicine, achievements of physics were always put to good use. Let us illustrate this with several examples from the twentieth century: discovery of electromagnetic waves — microwave therapy, discovery of X-rays — X-ray diagnostics and therapy, discovery of radioactivity — radiodiagnosis and radiotherapy, invention of lasers — laser therapy and laser surgery, etc.

From the textbook, you can see that in almost any branches of physics there are medical applications of physical knowledge and equipment, and medical devices, in fact, are entirely based on physical laws, regularities, physical phenomena, physical properties of materials, etc.

That is why physicomathematical, and biophysical knowledge is an essential element of higher medical education. They contribute to detailed study of the human body. This is important for formation of medicine as an exact science.

The reader of the textbook will work as a physician in the 21st century; medicine will make full use of concepts of exact sciences; computers and medical equipment will be used in medical practice much wider.

Learning this course is not easy, but the time and effort spent will be rewarded when you do subsequent courses, and in practical activity of a physician, who is the major figure in the process of therapy.